The genetic epidemiology of absolute pitch

by

Elizabeth Theusch

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

[redacted]

Acknowledgments

I could not have done the work described in this dissertation alone. Thanks to my thesis advisor, Dr. Jane Gitschier, for initiating the genetics of absolute pitch project and for guiding my research endeavors. Her eternal optimism and creativity added extra flavor to the project and helped me to deal with risks and failures as they came my way. Thank you also to the other two members of my thesis committee, Dr. Bob Nussbaum and Dr. Steve Hamilton, who freely shared their ideas about the project with me, giving me the advice I needed to move forward.

Thank you also to former Gitschier lab members who worked on the absolute pitch project: Siamak Baharloo, Barbara Levinson, E. Alexandra Athos, Amy Kistler, and Jason Zemansky, and Hernan Consengco. I thank Jon Woo, Elaine Carlson, and the UCSF Genomics Core Facility for their assistance with genotyping and Sanger sequencing and their advice for the project. Thank you to Ian McCulloch, Androuw Carrasco, and Yuri Cheung, three undergraduate summer interns who assisted with candidate gene sequencing. Thank you to Anna Need and David Goldstein for providing genotype data for individuals with four Ashkenazi Jewish grandparents. I would especially like to thank Clement Chu at the UCSF Center for Advanced Technology for teaching me a lot about next-generation sequencing and for sequencing our first round of samples. Thank you also to Leath Tonkin at the UC Berkeley/QB3 genomics sequencing facility for sequencing subsequent rounds of samples. I thank Dr. Neil Risch for giving us advice about analysis methods and data interpretation. I acknowledge Aaron Calhoun for website development.

Chapter IV describes published work with two co-authors. Dr. Analabha Basu taught me a lot about linkage analysis methods and gave me advice on the interpretation of our linkage results, though I was the one that actually conducted the experiments and the analysis. Dr. Jane Gitschier's role was to supervise the project as my thesis advisor.

Our study would not have been possible without our study participants and their families, so we thank them for their participation and interest in the study. This material is based upon work supported under a National Science Foundation (NSF) Graduate Research Fellowship. The National Science Foundation and the National Association of Music Merchants (NAMM) as well as the UCSF Academic Senate, the UCSF Research Evaluation and Allocation Committee, and the UCSF School of Medicine also provided funding for this project. Dr. Gitschier was previously an Investigator with HHMI who funded part of this research.

I would like to thank my family for their unconditional love and support throughout my life: my parents, Lee Theusch and Sandy Kelly; my siblings, Cassie and Jeff; and my late grandmother, Marjorie Kelly. I would also like to thank my California family, the Melcons, for providing me with a weekend refuge and for their conversation and companionship as I journeyed through graduate school. Finally, I would especially like to thank my friend and colleague, Dr. Yien-Ming Kuo, and my husband, Joshua Melcon, for giving me the encouragement, advice, and laughter that I needed on a daily basis to survive graduate school.

The Genetic Epidemiology of Absolute Pitch

Elizabeth Theusch

Absolute pitch (AP), also known as perfect pitch, is a rare pitch-naming ability with unknown etiology. Some scientists maintain that its manifestation depends solely on environmental factors, while others suggest that genetic factors contribute to it. We hypothesized that certain genetic variants predispose individuals with sufficient musical training to develop absolute pitch. We sought to identify those variants and to learn more about the etiology of absolute pitch using survey and pitch-naming test data from our participants. Our survey and test data agreed with previous observations that pitch-naming ability correlates with an early age of musical training onset, and the data exhibited the accuracy and precision of pitch-naming by AP possessors. Our AP twin study indicated that genetic factors contribute to absolute pitch's etiology, but our segregation analysis revealed that it was not inherited in a simple Mendelian fashion.

After collecting DNA samples from informative individuals, we conducted linkage analyses on multiplex absolute pitch families genotyped with microsatellite and single nucleotide polymorphism (SNP) markers and found a region of significant linkage in families of European descent on chromosome 8q24.21, along with suggestive linkage regions on 7q22.3, 8q21.11, and 9p21.3. There was evidence for genetic heterogeneity both within and between populations of different ancestry. In parallel with the linkage study, we attempted to discover genetic variants that were associated with AP in the Ashkenazi Jewish population using a genome-wide association study, but no variants were conclusively associated with AP. We then searched for AP-predisposing genetic variants by first Sanger sequencing candidate genes in eight AP individuals and

subsequently conducting targeted next-generation sequencing to sequence almost all genes in the four candidate linkage regions in ten AP individuals. Although a number of candidate AP-predisposing variants emerged from these data, including many novel SNPs, limited follow-up analysis did not conclusively support the association of the variants with AP.

Overall, our study of the genetic epidemiology of absolute pitch indicated that it is a complex trait that is genetically heterogeneous, with environmental, epigenetic, and stochastic factors also perhaps contributing to its genesis.

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION

What is absolute pitch?

The rare musical ability absolute pitch (AP), also called perfect pitch, has been a subject of human fascination and scientific study for decades if not centuries.[1] Though a variety of definitions have been applied to absolute pitch, in our study we defined it as the ability to instantaneously identify and label tones with their musical note names, without the use of an external reference tone. AP is sometimes considered to be the auditory equivalent of labeling visible light frequencies with color names without referencing a rainbow or color wheel, an ability most humans possess.[2] Since AP requires the labeling of tones with musical note names, AP possessors by our definition are musically trained because they need to have learned the names of the musical notes in order to use them as labels. It is important to distinguish absolute pitch from relative pitch (RP), the learned ability to judge intervals between pitches. Individuals with good relative pitch, which is common in trained musicians, have the ability to name pitches after being given an external reference tone.

Additional definitions of absolute pitch exist, however. Some consider the ability to verbally produce a tone on command without a reference as an alternate or extra requirement for AP possession. Others do not require a verbal label or verbal reproduction of the tone and consider the reproduction of notes on an oscillator or other instrument to be sufficient evidence of AP possession.[3] Less rigorous definitions of absolute pitch include the ability to name the key in which music is played,[4] the ability to always begin singing popular songs on the same note,[5] the ability to name notes played on one instrument but not others,[6] the ability to consistently speak words in tonal languages with the same combination of frequencies,[7] and the ability to mentally calculate the interval and name notes after comparison to a single internalized reference note.[6] Though many of these definitions and abilities overlap to some degree, some may rely on additional cues, such as instrument timbre or proprioceptive cues from the vocal apparatus, and others are not as instantaneous and effortless as pitch-naming by AP possessors is typically thought to be. Due to these and other considerations in our study, we chose a stringent definition that could be easily assayed using an online test.

Testing for absolute pitch

To classify people as AP possessors in our study, we used a test that was developed previously.[8] Since AP is rare in the general population, the test was made available online to facilitate recruitment of participants into our genetics of absolute pitch study. The test consisted of a series of 40 pure (sinusoidal) tones and 40 digitized piano tones presented in 10-tone blocks. Each tone was presented for one second followed by a three second break, so participants only had four seconds to identify the tone by clicking on the appropriate key on the virtual piano keyboard (Figure 1) before the next tone was presented. Participants were allowed to take breaks between the 10-tone blocks if they chose to.

Tests were automatically scored and entered into a database after each attempt by participants. Participants were given a point for each note they correctly identified and 0.75 points for each semitone error. Those over the age of 45 were given full credit for semitone errors because pitch perception can shift with age.[9,10] Participants who scored at or above the threshold of 24.5 on the pure tone test were classified as AP-1 and were considered to have absolute pitch.[8] The piano tone test score was not used for this classification, because the piano tones may evoke additional cues for some participants.

No test is perfectly designed, and ours was no exception. First, it relied on the tones being reliably reproduced on the computers on which it was taken. However, since the test spanned more than six octaves, some computer speakers, especially laptops, failed to produce the notes at an audible volume, especially in the lowest octaves. Conversely, the highest notes were so high and sometimes so shrill that they were painful to listen to. Thus, four tones from each test in extreme octaves were removed from scoring for these reasons, so the maximum score for both the pure tone and piano tone sections was 36. Second, for those who were not adept with a computer mouse or their laptop computer's alternate pointing device, actually moving the mouse cursor and clicking on the appropriate virtual piano key was a time-consuming and sometimes impossible step in the time allotted. Third, participants who were not familiar with the Western or solfege scale would have had difficulty demonstrating their pitch-naming abilities on our test, and may have been false negatives. Fourth, participants who were familiar with the piano had an advantage in knowing the layout of the keyboard and also in identifying the piano tones, with their characteristic timbre. Finally, since the same sequence of tones was played on every test attempt, it is possible that a subset of participants could have cheated on the test by attempting it multiple times and using recording devices, tuners, or instruments to assist them in determining the correct answers to enter on later attempts.

Other methods have been used to test for absolute pitch. Some versions of pitch-naming tests attempt to destroy short term memory for pitches to deter those with good relative pitch but not absolute pitch by interspersing the tones with unrelated cognitive tasks like reading aloud[11] or with sounds meant to disrupt pitch memory, such as bursts of white noise[12] or glissandos.[13] There are also pitch memory tests for people who do not know musical notation, such as those that test for the ability to produce popular songs in the correct key,[5] the ability to identify the picture that was associated with a tone in training trials,[14] or the ability to reproduce a sounded note on a digital sine-wave function generator after a series of inter-stimulus distracting tones.[15]

Even infants have been tested to determine if they rely more on absolute pitch cues or relative pitch cues. This first involved giving the infants pre-test stimuli of a series of concatenated three-tone "words." They were then tested by presenting some of these tone words in isolation mixed in with three-tone "non-words" which had the same interval structure but different pitches than the words. Since infants respond to novelty, they were assumed to have remembered a previous stimulus like the current stimulus if they lost interest in it quicker than they did to the novel stimuli. They did indeed lose interest in "words" quicker than "non-words," indicating that they used absolute pitch rather than relative pitch cues during the task.[16]

Absolute pitch in animals

Though animals cannot be tested for absolute pitch in the same manner that we test humans for it, a small number of animal species have been assayed for AP using different testing procedures. One method was to use a training period in which animals were given a food reward for tones of some frequencies but penalized if they sought food after being presented with tones of other frequencies by the insertion of a rest period with the lights off before the next tone was presented. During the testing period, if the animals preferentially responded by seeking food after rewarded tones but not non-rewarded tones, it was evident that the animals could remember which tones were rewarded. Humans were tested in a similar fashion, except they indicated whether they thought a tone was a rewarded tone by pressing a button, not by retrieving a food reward. Under this kind of testing protocol, bird species such as zebra finches, white-throated sparrows, and budgerigars exhibited greater AP ability than (non-AP possessing) humans and rats.[17,18] It would be interesting to see how other mammals, such as non-human primates, and humans with absolute pitch would perform on this sort of test.

Other characteristics of absolute pitch

Based on data from our online test, most participants in our study either exceeded our threshold for absolute pitch ability or tested within the range of random guessing, with relatively few participants falling in between.[9] This bimodal distribution of pitch-naming ability differed from the distribution of many other complex traits, such as blood pressure or height, which approach a normal distribution. This implies that a smaller number of factors may be involved in the development of absolute pitch or that a number of factors have to combine in an all-or-none fashion to give rise to absolute pitch.

Systematic shifts in the perception of pitches by AP possessors have been reported with aging,[9,10] medication,[19-22] and hormonal fluctuations.[23] In the cases of medication and hormones, these changes appear to be reversible. It is likely that these changes in pitch perception are not unique to AP possessors but that AP possessors are unique in their ability to detect them. It is interesting to speculate what causes these changes; in the case of aging, for instance, it is possible that a physical increase in the elasticity in the basilar membrane of the cochlea[24] due to a decrease in extracellular matrix integrity with age could elevate the pitch map relative to established neuronal connections.[9]

Though AP is rare in the general population, occurring in approximately 1 in 10,000 individuals,[25] it is more common in certain sub-populations. The blind,[26] musicians,[8,27] and individuals of Asian descent[28] all have a higher reported prevalence of AP than does the general adult population. Absolute pitch has also been reported as more prevalent in individuals with autism[29] and Williams syndrome[30]. The link with autism is interesting given that individuals with autism or AP exhibit piecemeal information processing. AP individuals are also more likely to be socially eccentric and exhibit other language and behavioral features associated with autism than are control musicians.[31]

Neurological correlates of absolute pitch in humans

Pitch processing is required for the proper perception of sounds from music and language in the brain. After sounds activate the cochlea, the resulting signals travel up the auditory pathway to the auditory cortex, undergoing a limited amount of processing along the way.[32] Like the cochlea, the primary auditory cortex has a tonotopic organization, with low frequencies represented laterally and high frequencies represented medially.[33] The hemispheres are not symmetric, however, because neurons in the right primary auditory cortex are more sharply tuned to frequency,[33] while those in the left primary auditory cortex are more sensitive to the temporal characteristics of auditory input.[34] Tonal input has high spectral resolution while speech input has rapidly changing energy peaks,[35] so it makes sense that the human brain predominantly processes tonal sounds in the right hemisphere and speech sounds in the left hemisphere of the auditory cortex.[36]

Since absolute pitch is a cognitive ability, one might hypothesize that the brains of musicians with AP would differ from the brains of musicians with RP structurally and functionally. Indeed, magnetic resonance imaging (MRI) studies show enhanced left-larger-than-right asymmetry of the planum temporale, a temporal lobe region located just posterior to the primary auditory cortex (Figure 2), in AP musicians compared to RP musician controls and non-musicians.[37,38] Though AP musicians have slightly larger left planum temporales on average[37], it appears that the asymmetry difference can mainly be attributed to the smaller right planum temporales found in AP musicians[38]. Anecdotally, the anterior left temporal lobe does not appear to be necessary for AP absolute pitch ability, because AP remained intact in a 17-year old AP possessor after he had an anterior left temporal lobectomy.[39]

The brains of AP musicians also differ from those of RP musicians during higher pitch processing. One response to pitch perception is the P300 (P3), an electrophysiological reaction to a stimulus that can be measured using electroencephalography (EEG), which has a greater amplitude if a sensory stimulus is rare and a greater latency if it takes the brain longer to process the stimulus and update the working memory.[40] The average P300 of AP musicians in response to novel tone stimuli is significantly reduced in amplitude and latency compared to RP musician and non-musician controls.[41,42] Using positron emission tomography (PET) to measure cerebral blood flow, a more recent study showed that AP musicians but not RP musicians exhibit activation of their left posterior dorsolateral frontal region, a region implicated in conditional associative learning, in response to tones but not noise.[43]

Together, these studies suggest that AP and RP musicians rely on different neurological pathways to process tonal information. Though it is still unclear when these brain differences manifest themselves developmentally, it is possible that they have a genetic basis.

Why study absolute pitch?

It is important to address why a peculiar trait like absolute pitch, which has no direct impact on human health, merits scientific study. Absolute pitch is more amenable to study than other cognitive traits due to its all-or-none manifestation and the relative ease with which it can be assayed. Since absolute pitch is a complex trait, involving both genetic and environmental components, it could serve as a model for studying other traits with complex etiologies. Absolute pitch, like language, appears to develop during a critical period of childhood development, so the study of AP could give insights into brain development and plasticity. This would have implications for the study of learning, memory, and cognitive disorders, such as autism and Williams syndrome. Since music and language are in many ways parallel processes that involve the perception and production of sounds, the study of musical abilities like AP could give insights into language and language disorders. It would be interesting to study AP from an evolutionary perspective in other species as well, since it is not immediately obvious what selective advantage or disadvantage, if any, the possession of this trait would have for humans.

Nature versus nurture

The etiology of absolute pitch is complex. Although at least one scientist[2] felt that "the nature-nurture debate in this particular arena essentially ended with the death in 1957 of Bachem," current scientific literature suggests that a variety of environmental and genetic factors may play a role in the development of AP. Musical training during a critical period of childhood development[8,44-46] likely contributes to the acquisition of AP, but this training alone is insufficient since many people receive early musical training but do not develop AP. In fact, it is difficult to discern whether early musical training predisposes to AP or AP predisposes to early musical training. Other environmental factors have been suggested to influence whether an individual develops AP, including the type of musical training the individual received[47] and the individual's tone language fluency.[48] Development of AP may also be influenced by other cultural factors that have not yet been identified.

We and others hypothesize that the genetic makeup of the individual also contributes to the development of this ability.[8,28,49,50] Familial aggregation studies have estimated the sibling recurrence-risk ratio (λ_s) for absolute pitch to be between 7.8 and 15.1 after controlling for early musical training.[28,51] Twin observations, while limited, give further support to this hypothesis. Three pairs of monozygotic twins concordant for AP and one pair of dizygotic twins discordant for AP have been reported in the literature.[52] Together, these data suggest that a combination of environmental and genetic factors likely promote the genesis of AP.

Approaches to studying the genetic basis of complex traits

A number of different strategies can be employed to identify genetic variants that influence predisposition to a trait. Some rely on existing knowledge to choose candidate genes and/or other genomic regions as the starting point of the investigation, while others begin with an unbiased, genome-wide approach. Some use families as study participants, while others use unrelated individuals from the population(s) of interest. Some look for regions of the genome that are shared among related individuals with the trait (linkage), while others look for specific alleles or haplotypes that are enriched in individuals with the trait (association). Some test whether the genetic data support a certain model for the inheritance of the trait (parametric), while others can detect predisposing genetic variants that are inherited in a variety of different ways (non-parametric).

Regardless of its philosophy, each strategy employs genetic and physical maps of the human genome containing the positions of polymorphic sites (genetic markers) that are assayed in study participants.[53-57] Commonly used genetic markers include microsatellites (short tandem repeats) and single nucleotide polymorphisms (SNPs). If an allele of a polymorphic genetic marker tracks with the trait of interest within a family, it is linked to the trait, and if it is enriched among unrelated individuals with the trait, it is associated with the trait. Both methodologies allow the estimation of the genomic positions of candidate regions that may contain genetic variants that predispose to the trait of interest, because they would be physically close to linked or associated genetic markers within the genome.

Unlike simple Mendelian traits, complex traits may involve a variety of different genetic variants that could work in isolation or in concert to give rise to a trait. Thus, a parametric linkage study using a dominant or recessive model may not work as well for a complex trait as it would for a Mendelian trait involving one gene and an unambiguous pattern of inheritance, especially if large families containing many individuals affected with the trait are not available. An alternate strategy is to use an affected relative pair approach for family-based linkage studies of complex traits.[58] In this strategy, the identity-by-descent probability of marker alleles is calculated for pairs or groups of affected relatives, giving an estimate of the likelihood that that allele (and thus the neighboring genomic region) is shared between related individuals with the trait.

With the advent of large-scale SNP genotyping platforms, another strategy, whole-genome association, became a reality.[59] Genome-wide association studies (GWAS) possess more power to detect genetic variants with small effect sizes than do linkage studies. However, in order for association to be detected, the predisposing genetic variant needs to be common enough that its haplotype is adequately tagged by at least one SNP of the hundreds of thousands that are typically genotyped in each individual for a GWAS.

Only rarely would an actual predisposing variant be genotyped directly during the first pass of a linkage or an association study. Due to the small number of recombinations in a typical family linkage study, linkage studies have less precision than association studies because relatively large segments of the genome can be shared by affected family members, and thus the results are generally candidate regions containing many genes and many potential predisposing variants. Though association studies can involve individuals who are not closely related, there are still small segments of the genome (linkage disequilibrium [LD] blocks) that are often inherited as one unit throughout many generations. Consequently, SNP alleles within those LD blocks are correlated with one another, so a neutral genotyped SNP may show association with the trait of interest because it is in the same LD block as a predisposing genetic variant. Whether it is a relatively large region from a linkage study or an LD block from an association study, fine mapping with additional genetic markers and/or re-sequencing approaches would then be employed to discover candidate predisposing genetic variants.

Hypothesis and objectives

We hypothesized that certain genetic variants predispose individuals to develop absolute pitch, assuming sufficient early musical training. The main goals of our study were to identify those variants and to learn more about the etiology of absolute pitch. We studied online survey and test data from AP and non-AP study participants to better understand how nature and nurture contribute to the development of the trait (Chapters II and IX). We then employed whole genome linkage analyses (Chapters III-V) and a whole genome association study (Chapter VI) followed by candidate gene re-sequencing (Chapter VII) and targeted next-generation sequencing (Chapter VIII) to investigate the genetic basis of absolute pitch.

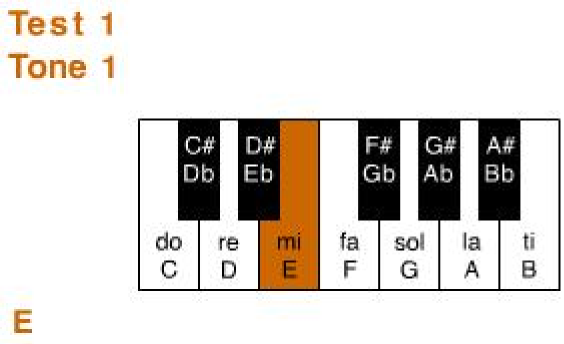


Figure 1. Virtual keyboard from our online absolute pitch test (http://perfectpitch.ucsf.edu).

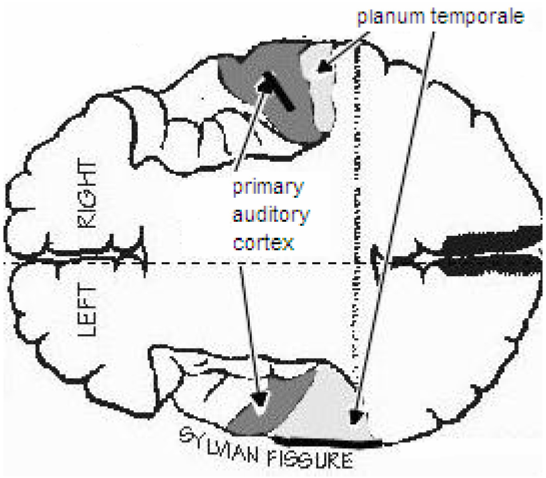


Figure 2. Position of planum temporale in the human brain.

II. TWIN STUDY AND SEGREGATION ANALYSIS

Most scientists agree that early musical training is important for the acquisition of AP, but some question whether there is a genetic predisposition for the development of AP. Though it has been shown that AP aggregates in families after controlling for age of musical training onset,[28,51] this aggregation may occur because other unknown environmental factors that influence AP aggregate in families. Twin and adoption studies are two ways to determine the relative contributions of genetic and environmental factors to a trait's etiology. To our knowledge, no adoption study has been conducted for AP. Three pairs of monozygotic twins concordant for AP and one pair of dizygotic twins discordant for AP have been reported in the literature,[52] though a larger sample size would be necessary to draw conclusions from twin data.

If the etiology of AP indeed has a genetic component, investigating the pattern of inheritance of the trait could lead to a better understanding of how many genetic variants may be involved and in what manner they interact with one another. Since pitch-naming ability appears to be a dichotomous trait rather than a continuous trait, it was proposed that absolute pitch could be influenced by only one or a few major genes.[9] One group conducted segregation analysis using a small sample of AP families with strong musical backgrounds and concluded that the inheritance pattern of AP was consistent with autosomal dominant inheritance with reduced penetrance, based on segregation ratios of .24 and .37, assuming single and complete ascertainment, respectively.[50]

We sought to further investigate the hypothesis that genetic factors are important for the acquisition of absolute pitch and to better elucidate the inheritance pattern of this trait. To this end, we conducted a twin study and a segregation analysis using data from a large population of absolute pitch possessors.

Subjects

Since absolute pitch is so rare in the general population, we used an online pitch-naming test[9] accompanied by an online survey as a recruitment tool to garner a large number of participants. The survey underwent two major revisions since it first appeared online in 2002, but all versions included questions about participants' contact information, demographics, musical training history, pitch-naming abilities, and family history. The most recent version of the survey (Appendix A), which was launched in February 2008, incorporated questions about the ethnicity and number of siblings of each participant.

Prior to February 2008, 16,504 participants, including some duplicates, took our online survey and/or test. Of these, 4,755 tested above our most stringent threshold for absolute pitch, being classified as AP-1.[8] Between February 2008 and March 2010, an additional 7,399 people participated in our test and revised survey, with 2,865 testing as AP-1 (38.7%). It should be noted that the frequency of absolute pitch possession in our study population is much higher than that of the general population and that of the musically trained population, probably because individuals with AP were more likely to find our website online and participate. This study was approved by the Committee on Human Research at the University of California, San Francisco. Participants were notified at the beginning of the survey that they were giving consent to participate in the survey and note-naming portion of the study by completing the survey and providing their contact information.

AP sibling report accuracy

Since we relied on the ability of AP-1 probands to accurately report whether their siblings have absolute pitch for both our twin study and segregation analysis, we first wanted to estimate the accuracy of those reports. To do this, we analyzed families who had participated in our study in which multiple siblings had taken our pitch-naming test. Out of 154 siblings of AP-1 probands who were reported by the AP-1 proband to possess absolute pitch, 133 of the siblings tested as AP-1. Of the 21 who did not, 10 scored above the AP-2, AP-3, or AP-4 thresholds,[8] indicating that they possess pitch-naming abilities that are well above average. Often, study participants who score above these less stringent thresholds for absolute pitch are able to exceed the more stringent AP-1 threshold if they take the test again. In addition, 1 sibling was only 5 years old and another 3 were over 50 years old, so the test scores for these 4 individuals may not have reflected their pitch-naming abilities at a different point in their lifetime.[9]

A conservatively high estimate for false positive reporting of sibling pitch-naming abilities by probands who are AP-1 is thus 21/154 or 13.6%. A lower estimate which assumes that all individuals who test as AP-2, AP-3, or AP-4 have absolute pitch is 7/150, or 4.7%. The false positive rate of 7.7% (1/13) from a previous study[51] falls between these upper and lower boundaries. As a side note, 7 individuals who were not reported to have absolute pitch by their AP-1 siblings were false negatives and were classified as AP-1 upon testing. Since we do not know the total number of siblings who were thought by the AP-1 probands not to have absolute pitch, we cannot estimate the false negative rate from our data.

Twin study

Of the individuals who tested AP-1 in our study to date, 30 probands reported being fraternal twins and 14 probands reported being identical twins, and each proband gave us additional information about the pitch-naming abilities of their twin. In some but not all cases these reports were validated if the second twin also entered our study and took our pitch-naming test.

Since zygosity information was based on participant reports, it may not have been completely accurate. Of the 14 monozygotic twin pairs, only 1 had been confirmed to be identical by genotyping done in our laboratory. Of the 30 dizygotic twin pairs, 12 pairs contained twins of the same sex (3 confirmed concordant, 2 reported concordant, and 7 reported discordant). We followed up with the 25 twin pairs who could potentially have been misclassified as fraternal or as identical due to concordance of gender. Of the monozygotic twin pairs, 4 reported being confident that they were identical based on their physical appearance, and 2 of these 4 reported being monoamniotic twins. Of the dizygotic twin pairs, 4 reported being confident that they were fraternal based on their physical appearance, 2 of these 4 reported being diamniotic and dichorionic twins, and 1 of these 2 reported having a different blood type than his twin. A fifth pair of twins who reported being fraternal was born with two placentas but reported that their appearance was "similar but not the same." The remainder of the twin pairs did not reply to our requests for additional information about their zygosity status.

Pairwise and casewise twin concordance rates were calculated under the assumption of single ascertainment (Table 1).[60] Pairwise concordance is the probability that both twins are affected given that at least one of the co-twins are affected, while casewise concordance is the probability that one co-twin is affected given that the other co-twin is affected.[61] Since casewise concordance is measured at the level of the individual rather than the twin pair, it is a more useful measure for comparison to risk rates among other pairs of relatives or estimates of population prevalence. In our study we assumed single ascertainment because we did not know of any instances in which each member of a twin pair entered our study independently of the other. Instead, the twin probands referred their co-twins to our study. Thus, concordant twin pairs were more likely to be ascertained than discordant twin pairs in our study, which was corrected for when single ascertainment was assumed.

As is evident in Table 1, the concordances of monozygotic twins were greater than the dizygotic twin concordances by 33-35%, depending on the type of concordance measured, supporting a role for genetic components in the etiology of absolute pitch. The standard errors of the concordance estimates were also calculated[60] and used to determine the significance of the differences between the concordances of monozygotic and dizygotic twins. Though the sample sizes for our twin study were not very large, they were still sufficient to achieve statistical significance with greater than 95% confidence. We predict that increases in sample size would increase the significance of the results. Interestingly, a larger twin study on another pitch perception ability, musical pitch perception tested with a distorted tunes test, indicated the involvement of genetics in the acquisition of that ability, with a monozygotic twin probandwise concordance of 75% and a dizygotic twin probandwise concordance of 57%.[62] Thus, genetic factors likely play a role in a variety of aspects of musical sound processing.

It is also important to note that the concordances of monozygotic twins were less than 100% in our study. Possible explanations for this include differences in environment between twins, including musical training, the influence of stochastic factors on penetrance of the trait, or epigenetic differences between twins.[63] For example, in one of our discordant monozygotic twin pairs, the co-twin reported not to have AP had no musical training, while the co-twin with AP had had training. Similarly, in another discordant monozygotic twin pair, the co-twin with AP had musical training on the guitar, while the co-twin reported not to have AP received training on the drums but no tonal instruments.

Simple segregation analysis

In the newest version of our online survey, participants were asked how many sisters and brothers they had, what their birth order was, whether they had family members with absolute pitch, and how many of each type of family member, including sisters and brothers, they had with absolute pitch. In addition, participants were asked about the ethnicity and country of origin of their ancestors. Of the 2,865 participants who tested with AP-1, 1,463 probands provided enough accurate information about their siblings to be included in our segregation analysis of absolute pitch families, as described below (Table 2). The largest number of participants reported East Asian ancestry (Table 3), and a substantial number also reported non-Ashkenazi-Jewish European ancestry (Table 4). Smaller numbers of participants reported Jewish ancestry, African ancestry, Hispanic ancestry, or mixed ancestry.

Participants were disqualified from segregation analysis if they provided inconsistent answers to survey questions or if we had reason to suspect that they were duplicates. If they did not disclose the numbers of brothers or sisters they had and they did not list their birth order as 1, they were excluded. Though this probably led to an underestimate of the number of AP probands that were only children in our dataset, this underestimate would not affect the segregation ratio. Participants were also excluded from segregation analysis if the birth order they listed was greater than the number of siblings they reported plus one, if they reported more AP siblings than they reported siblings, if they reported an unlikely large number of relatives with AP, or if they answered "No" or "Unknown" when asked whether they had family with AP and subsequently reported siblings with AP. In addition, if we could deduce that two or more participants were siblings from the same family, only the survey data from the initial proband was used.

On the rare instances in which we obtained multiple study participants from the same family, they often reported hearing about the study from a family member, so it is unlikely that two members of the same family were ascertained independently in our study. Thus, we assumed single selection and calculated the segregation ratio and its standard error using the method of Davie.[64,65] In sibships with an AP proband, the probability of a sibling of the proband being affected was p_D=(R-J)/(T-J), where R was the total number of AP siblings, T was the total number of siblings, and J was the number of sibships with only one proband. When all 1463 families were used (Table 2), p_D=0.089 with SEp_D=0.006. When only families with East Asian ancestry were included (Table 3), p_D= 0.096 with SEp_D=0.009. When only families with non-Jewish, European ancestry were included (Table 4), p_D= 0.078 with SEp_D=0.009. Since one would expect a segregation ratio of 0.25 for autosomal recessive inheritance and 0.5 for autosomal dominant inheritance, it appears that absolute pitch was not inherited in a simple Mendelian fashion in our families.

It is interesting that the segregation ratio using families of East Asian ancestry was slightly higher than the ratio using families of European ancestry. This indicates that AP-predisposing genetic and environmental factors may be operating somewhat differently in these two populations.

The segregation ratio estimates from our study were all substantially lower than those reported by Profita and Bidder over two decades ago,[50] probably due to differences in sample size and ascertainment criteria. While our study included 1463 families, their study included only 19 families. We included all probands who entered our study during a specific time frame that were not excluded based on inconsistencies in their survey data, as described above. The families from our study resided in many parts of the U.S. and the world, and the musical background of non-proband family members is unknown. In contrast, Profita and Bidder selected their AP probands from musical communities in large metropolitan areas and only chose probands who had musically educated families.

While their approach had the benefit of increasing the likelihood that family members had the necessary environmental influences to develop absolute pitch if they possessed predisposing genetic factors because they knew the musical note names to use on the pitch-naming test, it may also have enriched for families with multiple absolute pitch possessors. Though correlated, the cause-effect relationship of early musical training and absolute pitch possession is unclear. One could imagine that families with multiple absolute pitch possessors would have a higher degree of musical education than other families, and therefore there could have been ascertainment bias for families with a greater number of AP possessors.

Unfortunately, we did not have data from most of our probands about the musical training history of their nuclear family members, so we were unable to conduct a more complex segregation analysis incorporating environmental factors, such as musical training history, as covariates.

Absolute pitch relatives reported

Probands in our study were also asked to report whether they had other relatives with absolute pitch, in addition to siblings. Though these reports are potentially less accurate than the sibling reports, especially for more distant relatives, they still provided some information about how the risk for having absolute pitch varied with the relationship to the proband. Of the 2865 AP-1 probands, 422 (14.7%) reported that they had at least one family member with absolute pitch. Specifically, 195 of the 2189 siblings of probands (8.9%), 161 of the 5730 parents of probands (2.8%), and 90 of the 11460 grandparents of probands (0.8%) were reported to have absolute pitch. A variety of more distant relatives, such as uncles, aunts, and cousins, were also reported to have absolute pitch by the probands, but since we did not know how many total relatives each proband had in each of these categories, we could not determine the percentage of relatives reported to have AP in these categories.

The lower prevalence of absolute pitch possession in parents and grandparents of AP probands as compared to siblings indicates that absolute pitch is not inherited in a simple autosomal dominant fashion. Perhaps parents and grandparents of AP probands were less likely than siblings of AP probands to have sufficiently matched musical training to manifest absolute pitch if they were genetically predisposed to it due to spatial, temporal, and other differences in their childhoods. In addition, a combination of genetic factors may be necessary to develop absolute pitch, resulting in a more complex inheritance pattern as well. It is also worth noting that AP was reported to be more prevalent in dizygotic co-twins than it was in siblings of AP probands (45% vs. 8.9%), perhaps because twins have a greater shared environment, both pre- and post-natally, than do other siblings. Overall, the twin data, segregation analysis, and reports of AP relatives from our study suggest that absolute pitch is a complex trait with multiple genetic and environmental influences.

Table 1. Absolute pitch twin pairwise and casewise concordance rates.

	Monozygotic Twins				Dizygotic Twins			
Data Included	# of Pairs	# Conc	Pairwise Conc[a]	Casewise Conc[a]	# of Pairs	# Conc	Pairwise Conc[a]	Casewise Conc[a]
Confirmed	12	11	84.6%	91.7%	6	5	71.4%	83.3%
Reported	2	0	0%	0%	25	9	22.0%	36.0%
All	**14**	**11**	**64.7%**[b]	**78.6%**[c]	**31**	**14**	**29.2%**[b]	**45.2%**[c]

[a] Calculated under the assumption of single ascertainment[60]

[b] χ^2=4.56, 1 df, p=0.033

[c] χ^2=5.57, 1 df, p=0.018

Table 2. AP family distribution based on survey data from participants of all ethnicities.

#AP sibs		# of sibs per family											
per family	# families	1	2	3	4	5	6	7	8	9	10	11	12
1	1302	95	755	315	82	32	11	8		1	1		2
2	136		79	43	7	4	2	1					
3	17			14	3								
4	7				4	1	1			1			
5	1					1							
Total	1463	95	834	372	96	38	14	9	0	2	1	0	2

Table 3. AP family distribution based on survey data from participants who reported East Asian ancestry.

# AP sibs		# of sibs per family											
per family	# families	1	2	3	4	5	6	7	8	9	10	11	12
1	691	63	455	151	15	6							1
2	65		46	16		2	1						
3	10			9	1								
4	1									1			
5	1					1							
Total	768	63	501	176	16	9	1	0	0	1	0	0	1

Table 4. AP family distribution based on survey data from participants who reported non-Jewish European ancestry.

# AP sibs		# of sibs per family											
per family	# families	1	2	3	4	5	6	7	8	9	10	11	12
1	409	17	192	120	46	18	8	6		1	1		
2	44		17	21	3	2	1						
3	5			3	2								
4	4				2	1	1						
5	0												
Total	462	17	209	144	53	21	10	6	0	1	1	0	0

III. LINKAGE STUDY USING MICROSATELLITE MARKERS

Though it was hypothesized that genetic factors influence the development of AP, little was known about the molecular mechanism that gives rise to AP when our genetic study was begun over 5 years ago. To our knowledge, only one candidate gene approach had been attempted for AP and that was in the 7q11.23 region that is hemizygous in people with Williams syndrome. Williams syndrome is a cognitive disorder characterized by low IQ, poor conceptual and visual skills, elfin-like facial features, and other physical defects.[66] There have been reports that AP is more prevalent in individuals with Williams syndrome than the general population.[67] However, using 24 families with one or more affected sibling pairs, the study concluded that there was not significant allele sharing among sibling pairs at two microsatellite loci in the region.[68]

Lacking any obvious candidate genes to pursue, we chose a whole genome approach to the problem. As documented in Chapter II, AP is a complex trait with a pattern of inheritance that is not straightforward. This is likely due to the involvement of environmental factors in addition to multiple genetic factors.

When choosing whether to pursue a whole genome linkage strategy or a whole genome association study, the cost and power of the genetic analyses were considered. In general, association studies are well suited to identify common, low penetrance variants,[59] while linkage studies are better suited to identify rare, high penetrance variants.[69] Given the moderately high values of λ_s (7.8-15.1) estimated for AP,[28,51] the rarity of the trait and thus perhaps the rarity of the underlying genetic cause(s), and the expense of the technology to query a large number of markers at the time the project was started, a whole genome linkage strategy was pursued. This strategy was adopted because the causal variant(s) were assumed to have moderate to high effect sizes, because the causal variant(s) may not all reside on the same haplotype as certain alleles of queried, common markers, and because family-based analyses require fewer markers to be genotyped than population-based analyses due to fewer recombination events between individuals.

Though parametric LOD and HLOD statistics are typically more powerful than nonparametric linkage analysis,[70] a number of considerations discouraged this approach. AP has a complex and largely unknown pattern of inheritance. Due to the uncertainty in the estimates of penetrance and prevalence of AP, and the likelihood that multiple genes are involved, it would have been difficult to estimate the parameters and mode of inheritance for conventional, parametric linkage analysis, so a nonparametric affected relative pair approach was chosen. Power calculations indicated that affected relative pair linkage analysis had a good chance of success[58] given the sibling recurrence risk ratio and the assumption that genetic variants of major effect were playing a role in AP.

Prior efforts

Several steps had been taken towards determining the genetic basis of absolute pitch before I became involved in the project. In the first phase of the genome-wide linkage scan starting in 2004, the UCSF Genomics Core Facility (GCF) used the 400 marker ABI PRISM linkage mapping set v2.0-MD10 to genotype 38 families segregating absolute pitch (Figure 3). The markers in this set were di-nucleotide repeat microsatellites with an average spacing of about 10 centiMorgans (cM) throughout the genome. Alleles were separated on an ABI 3730xl capillary DNA analyzer and called using GeneMapper v3.5 software. Non-parametric linkage analysis was performed by Analabha Basu using the software package MERLIN (Multipoint Engine for Rapid Likelihood INference)[71] to locate regions of the genome which were co-inherited with AP in our families. Though there was a slightly promising region on chromosome 7 after this analysis, with a maximum nonparametric multipoint LOD score of 1.43 at marker D7S640, the LOD score diminished when genotype data from 9 additional markers were added to fine map the region.

In the second phase of the genome-wide scan, Elaine Carlson and Jon Woo of the UCSF GCF designed primers to query 376 additional markers that were used by the Center for Inherited Disease Research (CIDR). These markers were primarily tri-nucleotide and tetra-nucleotide repeat microsatellites, though some di-nucleotide repeats were also used to achieve an average spacing of about 10 cM in this CIDR set. The CIDR set was used to genotype 42 families, 36 of which had already been genotyped using the ABI marker set (Figure 3). (Families 85 and 5957 were genotyped with the ABI marker set but not the CIDR marker set.) When combined, these ABI and CIDR sets produced genotype data with an average spacing around 5 cM. The raw CIDR marker genotype data had been generated in 2006 before I began working on the project, but I played a major role in its analysis.

Further recruitment and genotyping

As we recruited additional multiplex absolute pitch families into our study through our website (http://perfectpitch.ucsf.edu) and collected their DNA in the form of blood or saliva samples, we genotyped them with a combined set of 700 di-, tri-, and tetra-nucleotide repeat microsatellite markers. We chose this combined set from the 776 markers in the ABI and CIDR sets discussed previously by eliminating 76 markers that performed poorly, were duplicates, or were located in close proximity to other markers. Since we had used these markers to genotype our previous absolute pitch families, it was relatively straightforward to combine the data from the various stages of genotyping for analysis. In total, DNA samples from 10 additional multiplex AP families (3957, 4404, 6057, 6734, 7734, 10435, 10644, 11155, 12125, and 12223) were genotyped with this combined microsatellite marker set (Figure 3). Also, 5 families that had only previously been genotyped with the CIDR markers (7701, 8133, 8141, 8210, and 9164) were genotyped with the ABI marker set. When genotyping was completed, only family 7959 was genotyped with the CIDR marker set but not the ABI marker set.

Following microsatellite genotyping using the ABI 3730xl DNA analyzer, the allele calls were made using GeneMapper v4.0. Since the allele-calling algorithms were not perfect, the microsatellite amplification reactions were not always optimal, and the pedigree data was not completely correct, errors needed to be detected and eliminated from the dataset. First, microsatellite genotypes that did not follow Mendelian segregation within families were detected with the assistance of PedCheck.[72] We then went back to the microsatellite traces in GeneMapper, identifying and correcting entries for any allele peaks which may have been miscalled. In the event that there were no obvious calling errors, we nullified the data for those family and marker combinations in the database.

Following the resolution of Mendelian inconsistencies, the genotyping data were analyzed for another class of errors, those in which the observed data could only be explained by double recombinations in close proximity. These errors were detected using the --error option in Merlin,[71] and due to their abundance and their sporadic nature, the majority were not investigated further and were simply eliminated from the dataset using the Pedwipe program in the Merlin package. Though we originally ran this analysis using the marker positions on the Marshfield genetic map, we later used the deCODE genetic map positions[57] in our final analysis due to its increased accuracy and correspondence with the physical map of the human genome. For those markers which were not on the original deCODE map, we used interpolated genetic distances.[73]

Nonparametric multipoint linkage analysis

The majority of our families (36) were of European descent (Eu), but we also had 11 of East Asian descent (EAsian), 6 of Ashkenazi Jewish descent, and 1 (3613) of Indian descent (Figure 3). Since allele frequencies differed in these different populations, we subdivided the linkage analyses as follows: one using the families of East Asian descent only, one using the families of European, Ashkenazi Jewish, and Indian descent together (Eu/AJ/I), and one using only the families of European descent, excluding those of Ashkenazi Jewish and Indian descent. The frequencies of the various microsatellite alleles were estimated from the founders of each set of families.

We used the software package Merlin[71] for our non-parametric linkage analysis, which incorporates a modified Lander-Green algorithm that is less memory-intensive than the original Lander-Green algorithm[74] but can handle a large number of markers unlike the Elston-Stewart algorithm.[75] Merlin is faster than its predecessors, can handle loops and other pedigree complications, works well on datasets with a large numbers of markers but relatively small pedigrees like ours, and is easy to use and is well documented.

The initial linkage analyses were performed in Merlin[71] using Whittemore and Halpern S_{all} statistics[76] and Kong and Cox linear and exponential logarithm of the odds (LOD) scores.[77] We conducted nonparametric linkage analyses using both the exponential and linear models[77] because we had a relatively small number of families and were not sure whether our analyses should be optimized to detect a large increase in allele sharing among affected relatives in those families (exponential) or permit a lesser degree of allele sharing (linear). Our initial approach was to conduct multipoint linkage analysis. Though multipoint analysis utilizes more information and is theoretically more powerful than single point analysis, multipoint analysis is more sensitive to genotyping error.[78] Thus, we also conducted single point linkage analyses since microsatellite linkage data typically contain some errors.

Figures 4A and 4B display the nonparametric multipoint linear and exponential LOD scores, respectively, at the location of each marker assayed. Separate analyses were conducted on the Eu, Eu/AJ/I, and EAsian families. Regions which showed the most promising signs of linkage are summarized in Table 5. Theoretically, for an affected sibling pair linkage study, LOD scores of at least 2.2 are considered suggestive of linkage, while those which are at least 3.6 exhibit significant evidence for linkage.[79] The Eu and Eu/AJ/I LOD scores were quite correlated with one another, as were the linear and exponential LOD scores, so we did not correct for multiple testing.

Nonparametric multipoint linkage analysis after fine mapping

The candidate regions identified by the linkage analyses were quite broad, so it was necessary to perform finer mapping of the regions to shorten the list of potential candidate genes in those regions. Additional markers in those regions might also be more informative in our population than the ones in our initial marker set. Therefore, we genotyped 32 additional markers (Table 6) in and around some of the candidate regions in an effort to more precisely define sites of recombination in our families and to generate more accurate LOD scores. An additional 7 families had been added to the study by this time (Figure 5), and they were genotyped only for these markers.

As is evident in Figure 6, the additional markers did not drastically change the LOD scores in the candidate regions. In the regions where there were some changes, the additional markers generally led to slightly decreased maximum LOD scores for the region. Overall, no regions approached the threshold for genome-wide significance. While this result provides no compelling support for a genetic basis for AP, it is also likely that the study was underpowered and recruitment of additional families would be needed to achieve significance. It is also possible that there are many different factors that contribute to AP but that each only has a small effect size, making their linkage signals difficult to detect in our study.

Ultimately, as will be seen in Chapter V, the Eu linkage regions on chromosome 8 around 145 cM and chromosome 9 around 47 cM remained after the addition of SNP markers, while those on chromosomes 4 and 12 did not hold up well.

Nonparametric single point linkage analysis

The results from nonparametric single point linkage analysis for all of the markers in our three different populations are shown in Figure 7. For some regions, the nonparametric single point LOD scores were higher than the corresponding multipoint LOD scores (Table 7), which could indicate that some markers in the regions had errors in assumed position or allele calling. Sometimes the marker with the highest local single point LOD score was not the same as the marker with the highest multipoint LOD score in the region. Some of our most promising multipoint linkage regions did not have markers with high single point LOD scores (Tables 5 and 7). This could indicate that the markers in those regions were not very informative independently, perhaps due to low heterozygosity. Though the average microsatellite marker heterozygosity was 76.5%, some markers had heterozygosities at or below 50%.

Parametric linkage analysis

Though the prevalence, penetrance, and mode of inheritance of absolute pitch was difficult to determine from our existing family data, we decided to conduct parametric linkage analysis in Merlin with different models to determine which of the models best fit the data from some of our most promising nonparametric linkage regions. Both multipoint (Figure 8) and single point (Figure 9) parametric linkage analyses were conducted on the Eu (Figures 8A and 9A) and the EAsian families (Figures 8B and 9B) using four different models. Heterogeneity LOD scores (HLODs) are shown because we assume that there is locus heterogeneity in AP. The details of the AP prevalences and penetrances assumed in the four models are summarized in Table 8. Using both multipoint and single point parametric linkage analysis, we examined each promising nonparametric linkage region and decided which models fit best for the region (Table 9).

It was often difficult to determine whether a dominant, recessive, or mixed model worked best for some regions, but for other regions it was quite clear. For instance, AP predisposing genetic variants in the chromosome 9 linkage region are likely inherited in a dominant fashion in Eu families, while those in the chromosome 12 linkage region are more likely inherited in a recessive manner. This information could become useful as candidate variants are discovered and tested for their segregation within families.

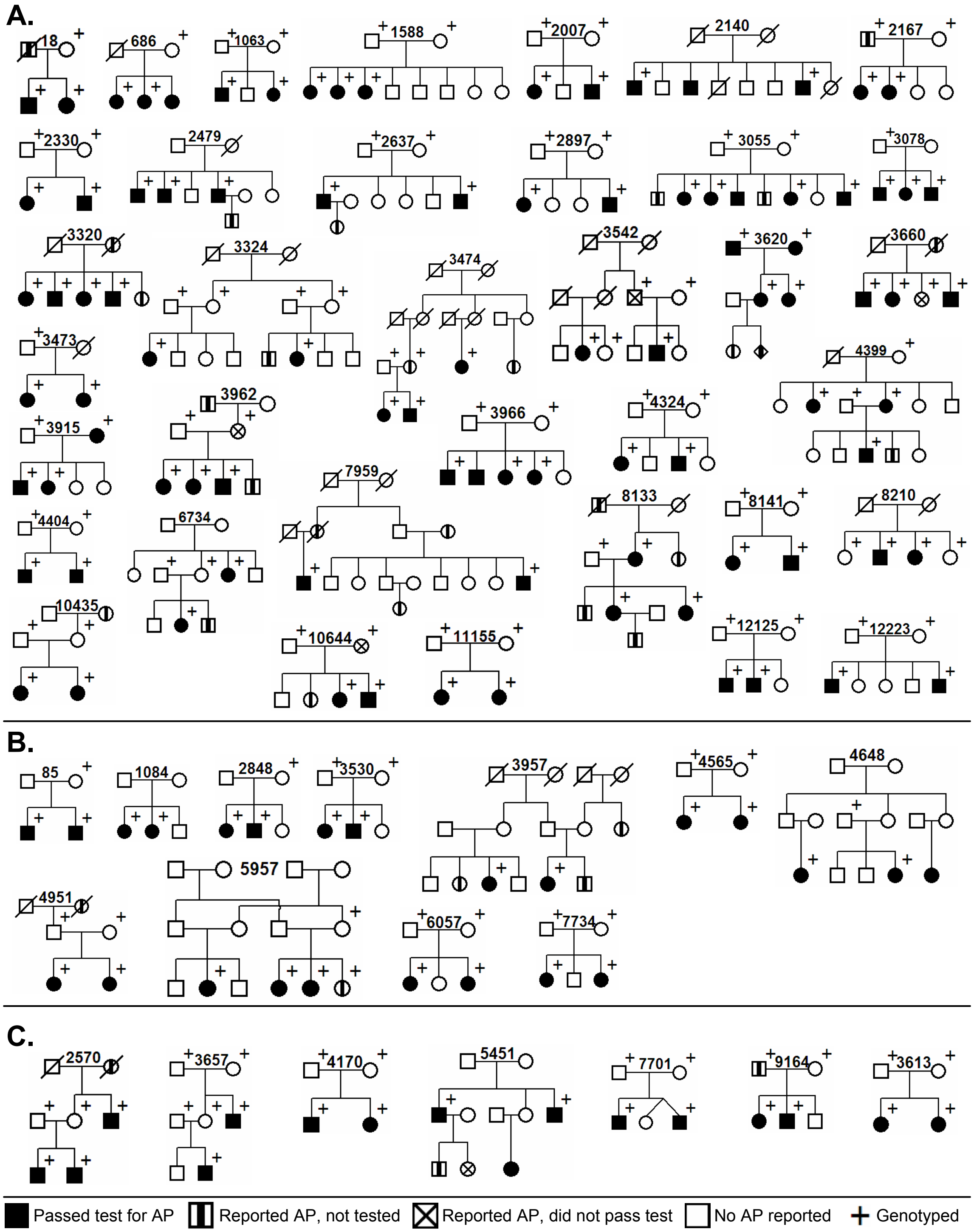


Figure 3. Families used for whole genome microsatellite linkage analysis. These include (A) 36 Eu families, (B) 11 EAsian families, and (C) 6 AJ families and 1 Indian family (3613).

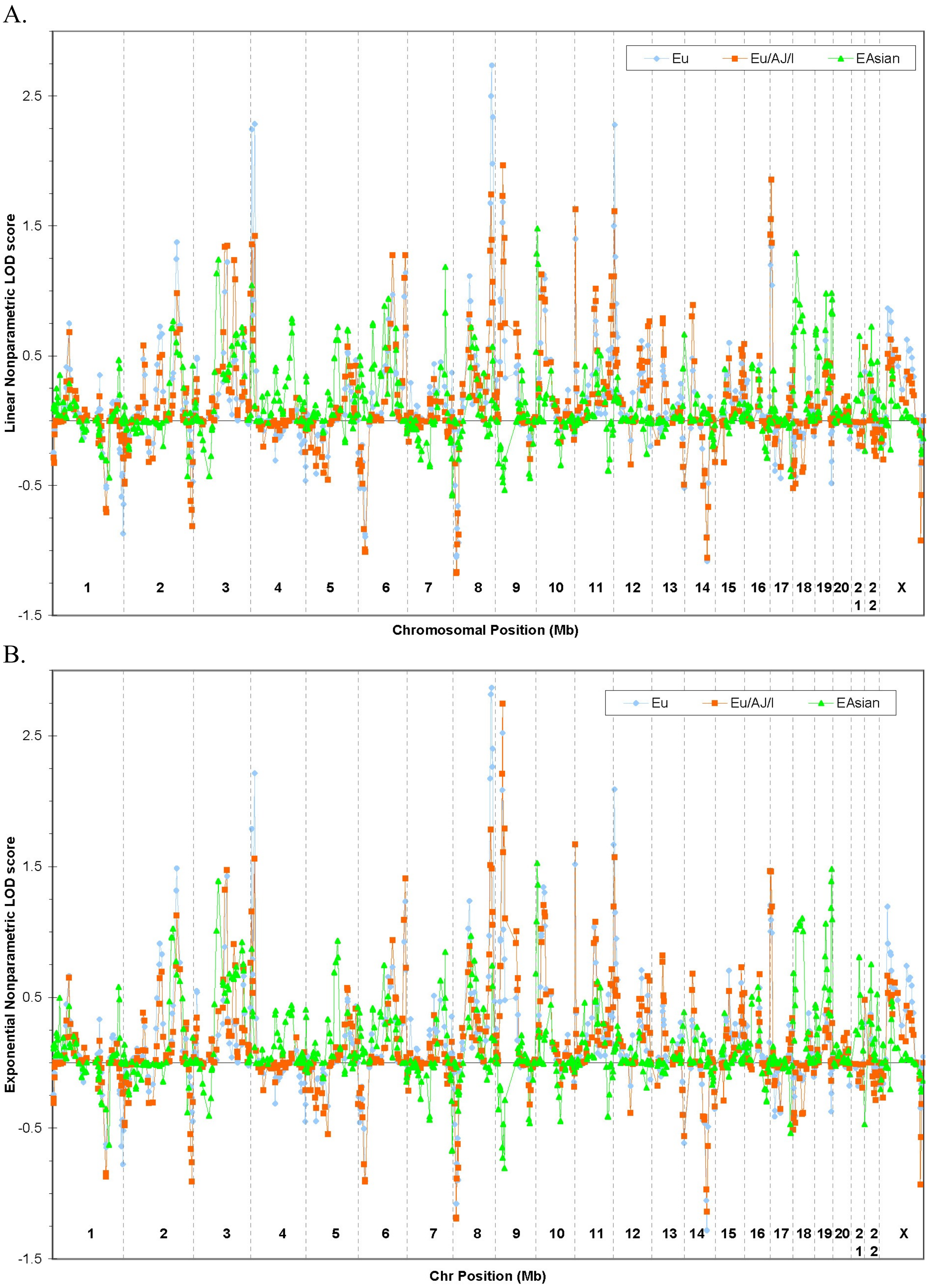


Figure 4. Linear (A) and exponential (B) multipoint nonparametric LOD scores at each microsatellite marker position across the genome.

Table 5. Most promising nonparametric multipoint linkage regions. Genetic distances are in deCODE cM.

Pop	Chr	cM	Marker	LinLOD	ExpLOD
Eu	2	189.15	D2S1391	1.376	1.488
Eu	3	122.74	D3S1278	1.222	1.427
Eu	4	26.71	D4S403	2.286	2.251
Eu	6	176.12	D6S1277	1.138	1.234
Eu	8	68.31	D8S285	1.115	1.237
Eu	8	145.26	D8S256	2.737	2.868
Eu	9	46.70	D9S1121	1.685	2.522
Eu	10	50.04	D10S197	1.126	1.343
Eu	11	1.26	D11S4046	1.401	1.517
Eu	12	8.43	D12S372	2.279	2.091
Eu	17	10.30	D17S1298	1.338	1.094
Eu	X	43.85	DXS9896	0.866	1.194
Eu/AJ/I	2	189.15	D2S1391	0.981	1.128
Eu/AJ/I	3	122.74	D3S1278	1.346	1.474
Eu/AJ/I	4	26.71	D4S403	1.423	1.561
Eu/AJ/I	6	176.12	D6S1277	1.277	1.408
Eu/AJ/I	8	139.79	D8S284	1.741	1.782
Eu/AJ/I	9	46.70	D9S1121	1.966	2.745
Eu/AJ/I	10	50.04	D10S197	1.015	1.207
Eu/AJ/I	11	1.26	D11S4046	1.629	1.671
Eu/AJ/I	12	8.43	D12S372	1.613	1.573
Eu/AJ/I	17	10.30	D17S1298	1.855	1.462
EAsian	2	177.98	D2S335	0.768	1.026
EAsian	3	108.71	D3S4529	1.241	1.389
EAsian	4	4.42	D4S412	1.043	0.872
EAsian	5	116.61	D5S2501	0.723	0.932
EAsian	7	137.75	D7S1804	1.184	0.847
EAsian	8	73.60	D8S260	0.721	0.970
EAsian	10	14.38	D10S591	1.481	1.527
EAsian	18	37.67	D18S542	1.291	1.021
EAsian	19	104.01	D19S418	0.935	1.481

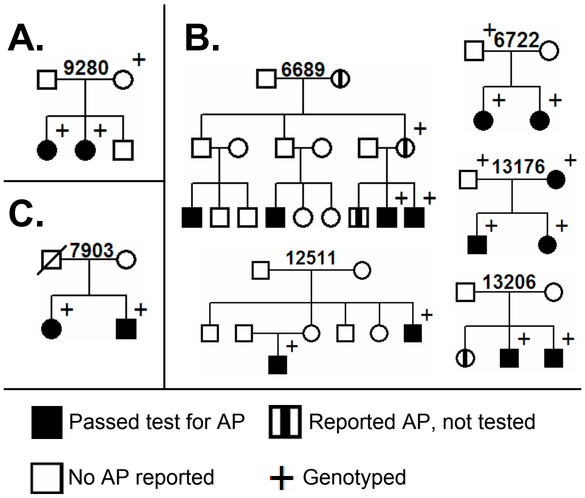
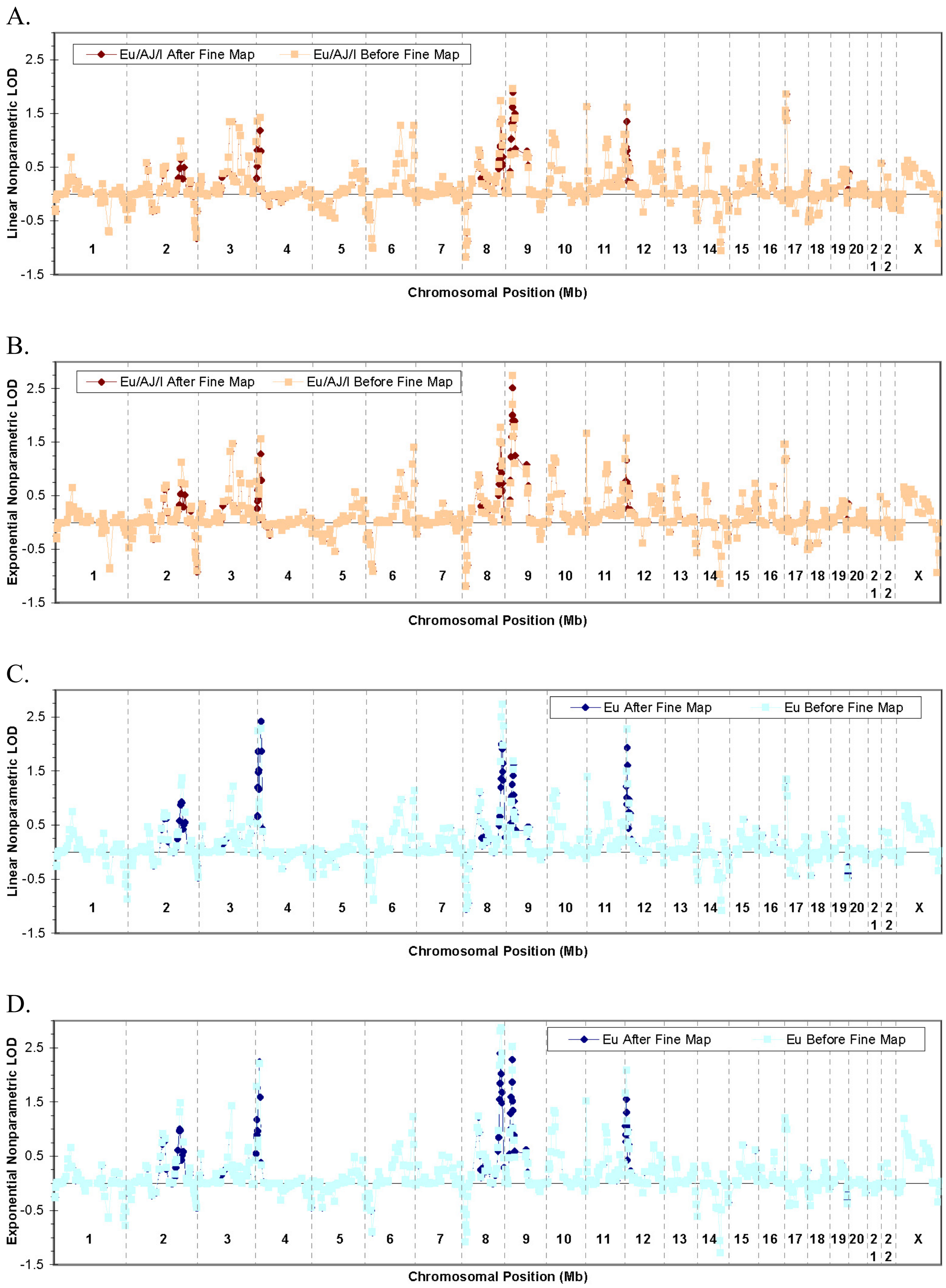


Figure 5. Families added for microsatellite fine mapping. These include 1 Eu family (A), 5 EAsian families (B), and 1 AJ family (C).

Table 6. Regions in which markers were added for fine mapping.

			Genotyped in:		
Chr	cM	# added	Eu	AJ/I	EAsian
2	189.15	4	x		
3	108.71	4			x
4	26.71	6	x	x	x
8	73.60	3			x
8	145.26	2	x	x	x
9	46.70	4	x		
12	8.43	4	x		
19	104.01	5			x


A.
Linear Nonparametric LOD
Eu/AJ/I After Fine Map
Eu/AJ/I Before Fine Map
2.5
1.5
0.5
-0.5
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosomal Position (Mb)
B.
Exponential Nonparametric LOD
Eu/AJ/I After Fine Map
Eu/AJ/I Before Fine Map
2.5
1.5
0.5
-0.5
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosomal Position (Mb)
C.
Linear Nonparametric LOD
Eu After Fine Map
Eu Before Fine Map
2.5
1.5
0.5
-0.5
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosomal Position (Mb)
D.
Exponential Nonparametric LOD
Eu After Fine Map
Eu Before Fine Map
2.5
1.5
0.5
-0.5
-1.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosomal Position (Mb)

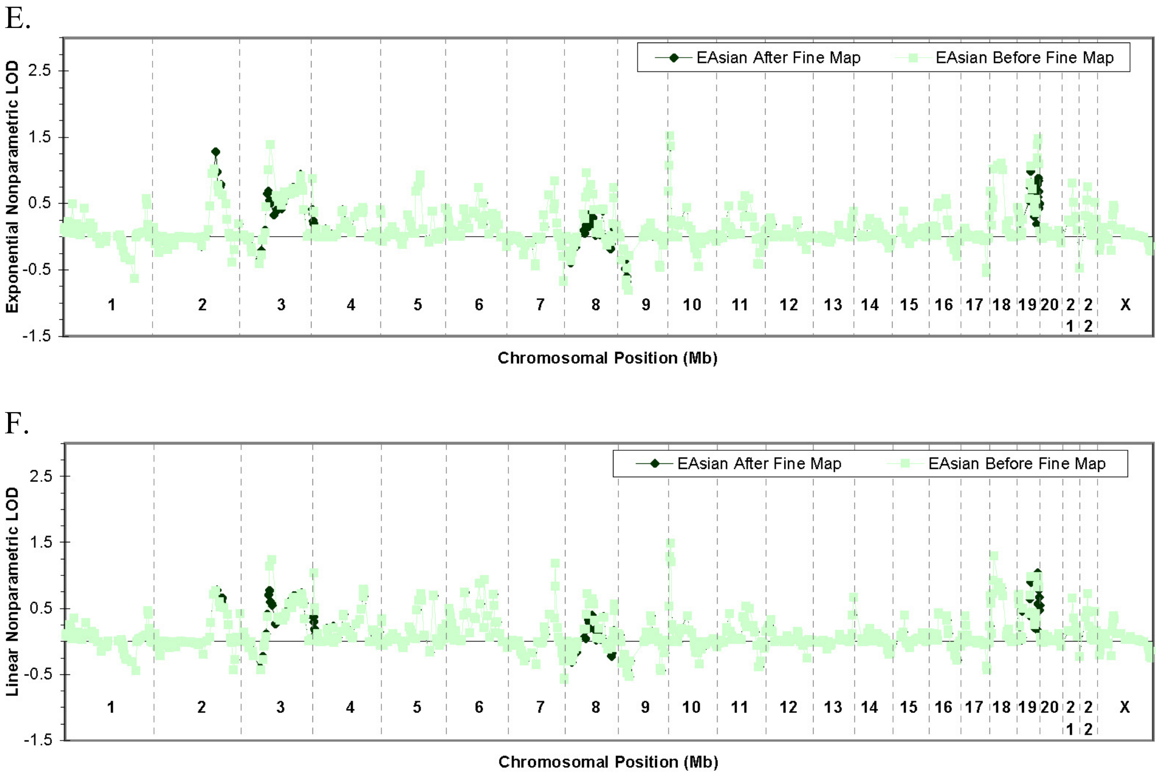


Figure 6. Nonparametric linear (A,C,E) and exponential (B,D,F) LOD scores before and after 32 markers were added for fine mapping. Results are shown for Eu/AJ/I families (A-B), Eu families (C-D), and EAsian families (E-F).

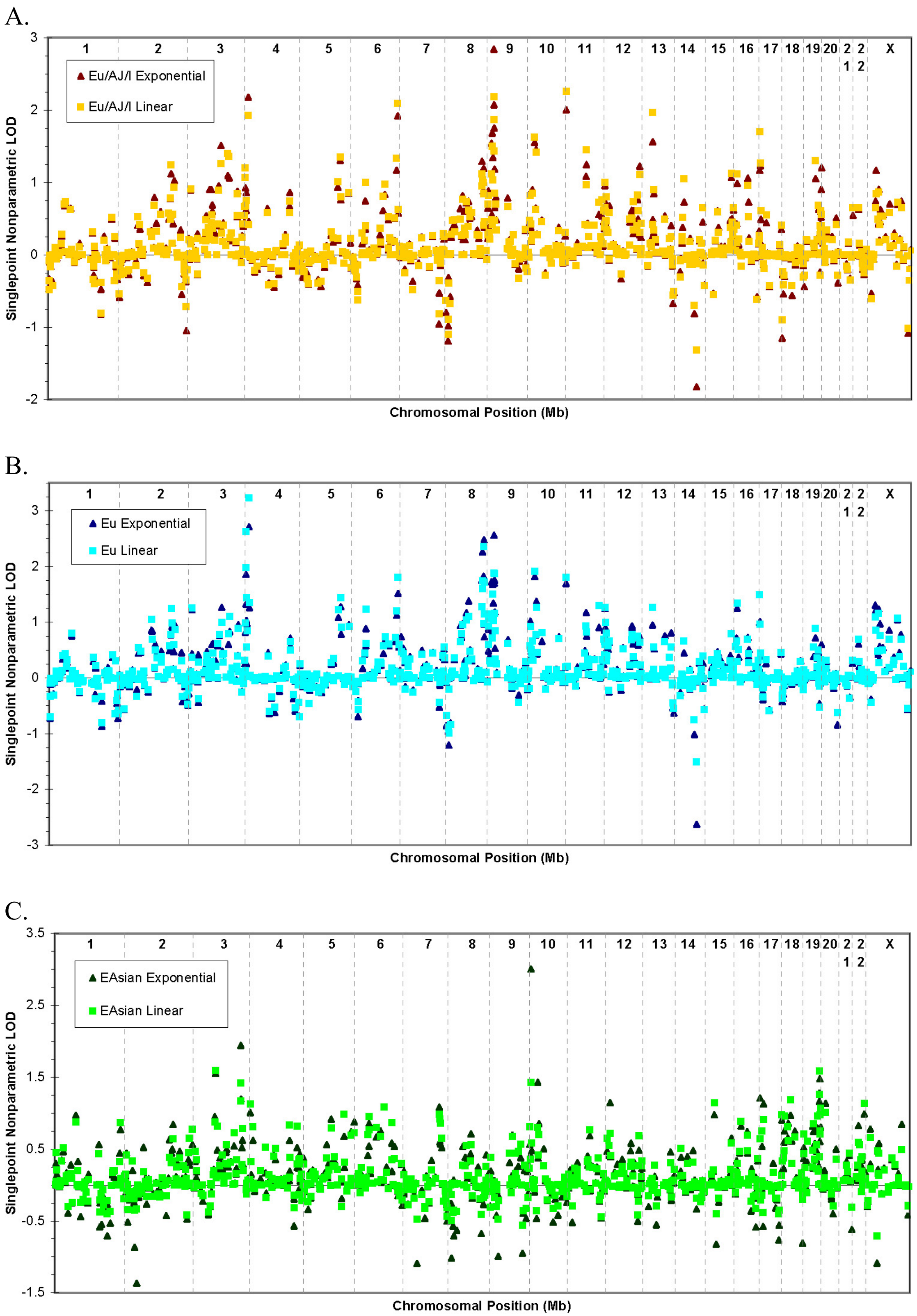


Figure 7. Nonparametric single point linkage results for Eu/AJ/I families (A), Eu families (B), and EAsian families (C).

Table 7. Comparison of most promising nonparametric singlepoint LOD scores to nearby multipoint LOD scores. Genetic distances are in deCODE cM.

Singlepoint linkage maxima						Local multipoint maxima		
Pop	Chr	cM	Marker	LinLOD	ExpLOD	LinLOD	ExpLOD	cM, marker
Eu	4	26.71	D4S403	3.229	2.713	2.419	2.240	same
Eu	6	176.12	D6S1277	1.809	1.519	1.138	1.234	same
Eu	8	148.14	D8S1108	1.722	2.479	1.995	2.403	139.79, D8S284
Eu	9	45.57	D9S171	1.876	2.562	1.617	2.287	46.70, D9S1121
Eu	10	50.04	D10S197	1.914	1.815	1.126	1.343	same
Eu	11	1.26	D11S4046	1.809	1.696	1.401	1.517	same
Eu/AJ/I	4	26.71	D4S403	1.928	2.117	1.180	1.279	same
Eu/AJ/I	6	176.12	D6S1277	2.095	1.920	1.277	1.408	same
Eu/AJ/I	9	45.57	D9S171	2.180	2.840	1.890	2.515	46.70, D9S1121
Eu/AJ/I	10	50.04	D10S197	1.627	1.554	1.015	1.207	same
Eu/AJ/I	11	1.26	D11S4046	2.262	2.002	1.629	1.671	same
Eu/AJ/I	13	38.54	D13S894	1.966	1.559	0.790	0.821	39.34, D13S218
EAsian	3	107.43	D3S3681	1.590	1.556	0.772	0.683	same
EAsian	3	168.52	D3S1763	1.419	1.942	0.734	0.938	same
EAsian	10	19.78	D10S189	1.429	3.006	1.481	1.527	14.38, D10S591
EAsian	10	57.66	D10S1426	0.805	1.433	0.387	0.444	70.07, D10S196
EAsian	19	96.54	D19S572	1.587	1.281	1.037	0.859	94.77, D19S589

Table 8. Models used for parametric linkage analysis. Each individual was categorized based on musical training initiation age, and this information was used as a covariate.

		Model			
Model parameters		RareDom	ComDom	ComRec	ComMix
	Prevalence	0.01	0.1	0.1	0.1
	Penetrance, 0 alleles	0	0	0	0
No musical training	Penetrance, 1 allele	0	0	0	0
	Penetrance, 2 alleles	0	0	0	0
Musical training after age 8	Penetrance, 1 allele	0.1	0.1	0	0.05
	Penetrance, 2 alleles	0.1	0.1	0.1	0.1
Musical training before age 9	Penetrance, 1 allele	0.9	0.9	0	0.45
	Penetrance, 2 alleles	0.9	0.9	0.9	0.9
Unknown amt of musical training	Penetrance, 1 allele	0.3	0.3	0	0.15
	Penetrance, 2 alleles	0.3	0.3	0.3	0.3

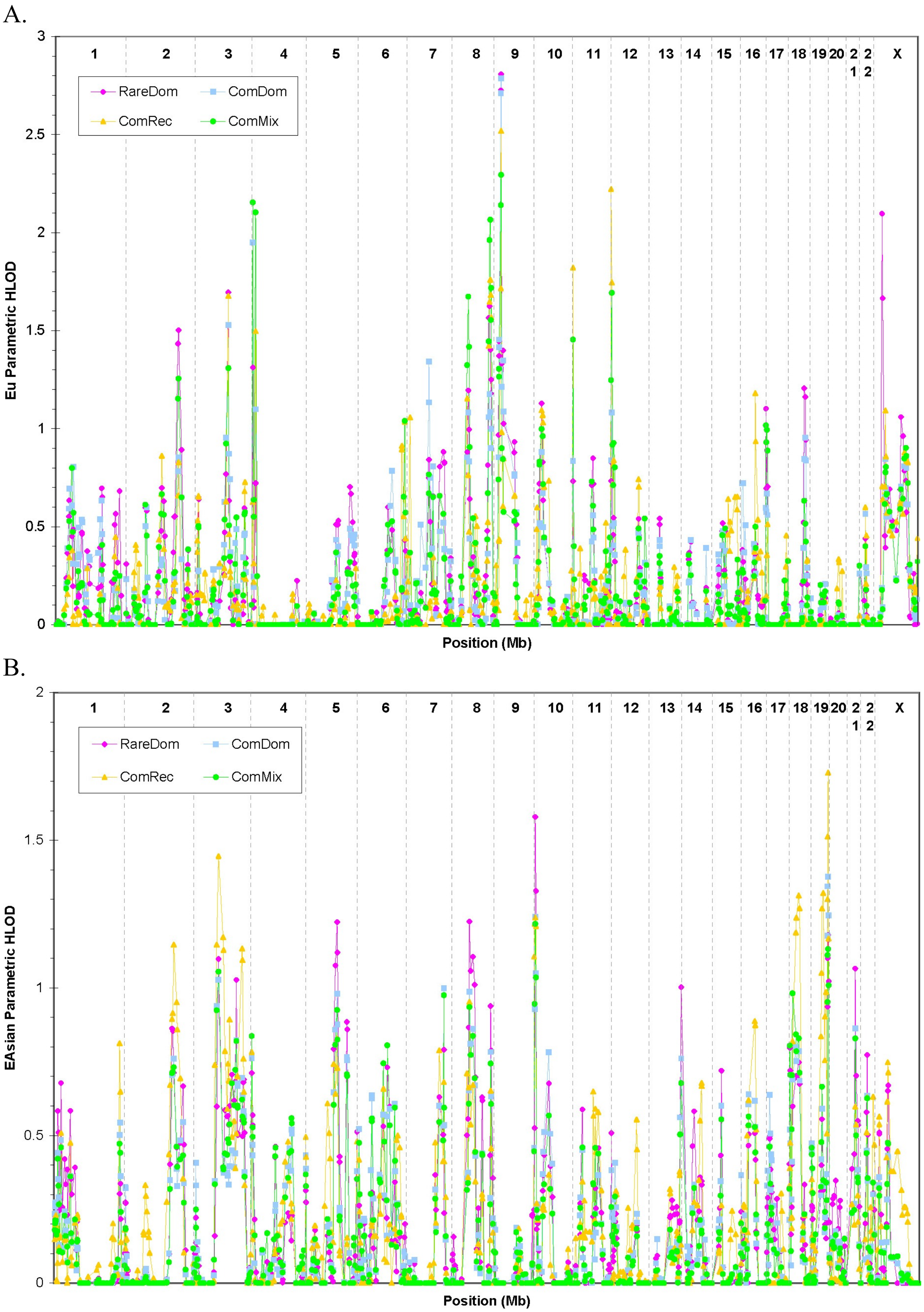


Figure 8. Multipoint parametric heterogeneity LOD scores from four different models using Eu (A) and EAsian (B) families.

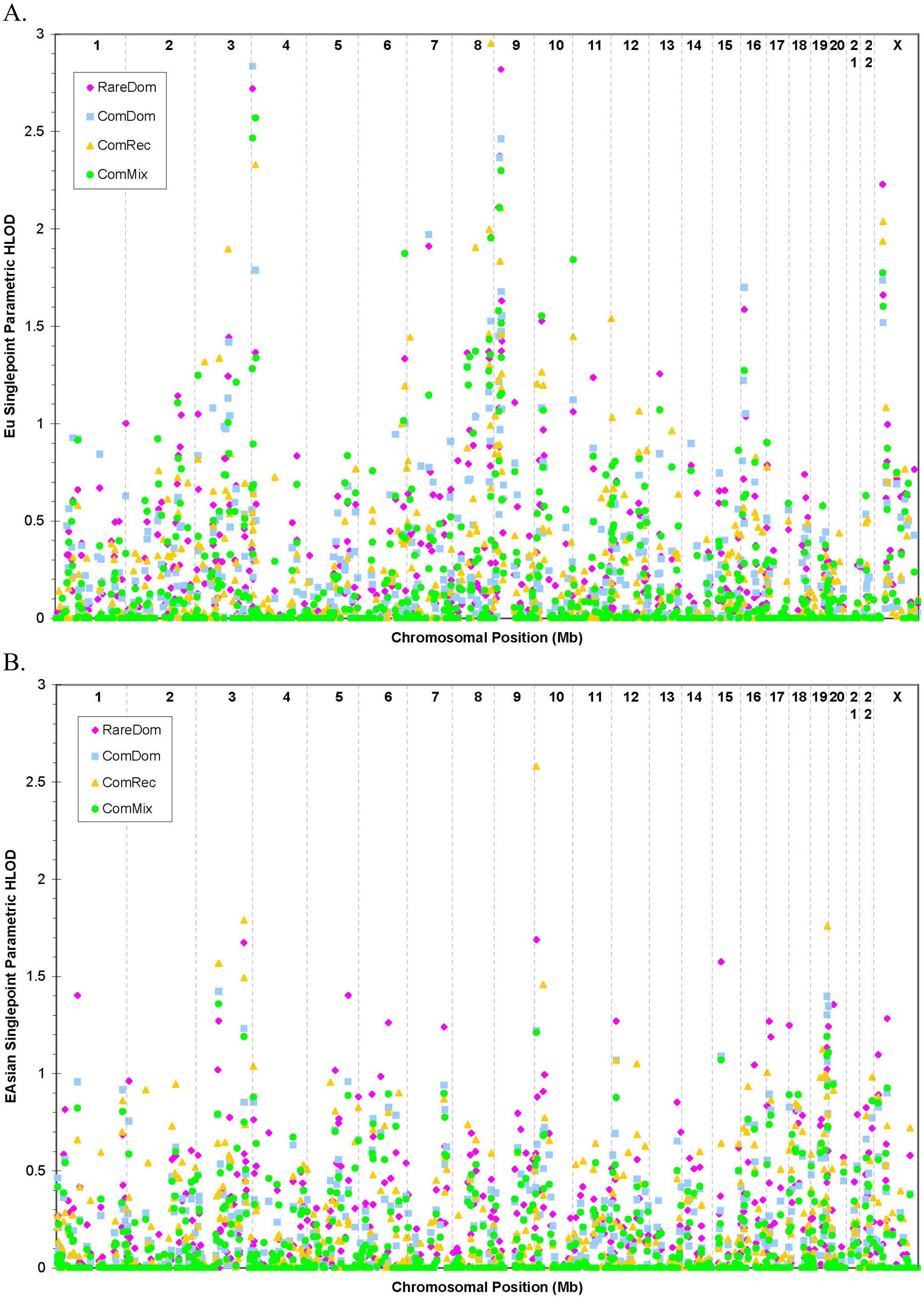


Figure 9. Single point parametric heterogeneity LOD scores from four different models using Eu (A) and EAsian (B) families.

Table 9. Best-fitting parametric linkage models from multipoint and singlepoint analysis for selected nonparametric linkage regions.

Pop	Chr	cM	Best model	Alternate model
Eu	3	122.74	ComRec	RareDom
Eu	4	4.42	ComMix	ComDom
Eu	8	145.26-148.14	ComRec	ComMix
Eu	9	45.57-46.70	RareDom	ComDom
Eu	11	1.26	ComRec	ComMix
Eu	12	0	ComRec	ComMix
Eu	X	43.85-46.21	RareDom	ComRec
EAsian	3	107.43-108.71	ComRec	RareDom
EAsian	5	116.61	RareDom	ComDom
EAsian	8	73.6	RareDom	ComDom
EAsian	10	14.38-19.78	ComRec	RareDom
EAsian	19	104.01	ComRec	ComDom

IV. LINKAGE STUDY USING SNP MARKERS

This chapter contains portions of a previously published manuscript[80] that have been modified from the original version.

As we recruited additional families, we switched from genotyping our families with microsatellites to genotyping them with single nucleotide polymorphisms (SNPs), due to the lower cost and greater data quality. Though SNPs are not as polymorphic as microsatellites are individually, their increased density compensates for the relative lack of heterogeneity. We conducted whole-genome, nonparametric linkage analyses on multiplex AP families genotyped with SNPs and successfully identified a region of significant linkage. Moreover, we found evidence for genetic heterogeneity both within and between populations of different ancestry.

Subjects and Methods

To facilitate the recruitment of individuals with AP, we employed our online pitch-naming test and survey, as described previously.[8,9] Participants who exceeded our threshold for AP on our pitch-naming test and who reported at least one relative with AP were asked to invite their relative(s) to also enter the study via the website. Study participants from families in which AP ability was documented in at least two family members who were not simply a parent-child relative pair and who resided in the United States or Canada were invited to contribute DNA samples to our linkage study. Participating family members were also encouraged to invite other family members who may be informative for our genetic analysis to contact us and provide a DNA sample, even if they did not possess AP. Participants who chose to donate mouthwash or saliva samples were given kits for self-collection of these samples. Blood samples were collected by a mobile phlebotomy service (ExamOne), and many of these were immortalized by Epstein Barr Virus (EBV) transformation.[81] DNA was extracted from mouthwash samples, whole blood, and lymphoblastoid cell lines with Gentra Puregene DNA purification kits (QIAGEN). Saliva samples were collected in Oragene DNA self-collection kits and purified according to the manufacturer's instructions (DNA Genotek). Written informed consent was obtained from all participants who contributed DNA samples to our study.

Overall, DNA samples from 73 families with at least one non-parent-child AP relative pair were collected for linkage analysis (Figure 10). These families included some of those who had been previously genotyped by microsatellite markers; however, some families studied previously did not have sufficient DNA remaining to be included in the SNP study. Nineteen families reported predominantly East Asian ancestry (E Asian), eight families reported being Ashkenazi Jewish (AJ), one family was Indian (I), and the remaining 45 families were predominantly of mixed European ancestry (Eu) (Table 10). Ancestry information reported by family probands correlated well with how the probands clustered on a multidimensional scaling plot generated using pairwise identity by state (IBS) distances calculated in Plink (Figure 11).[82] The distribution of AP relative pairs in the families is summarized in Table 11.

DNA samples from 281 individuals (indicated by the + signs in Figure 10) were genotyped with 6,090 SNPs on the Infinium HumanLinkage-12 BeadChip (Illumina) in the UCSF Genomics Core Facility. These SNPs were located at an average spacing of 0.58 cM (441 kb) throughout the human genome, and their genetic map positions have been estimated on the deCODE genetic map.[57] The Genotyping Module of Illumina's BeadStudio software was used to manually inspect SNP genotype calls on intensity plots. Once obvious errors were resolved, the genotype data were analyzed with Pedcheck to locate Mendelian inconsistencies.[72] These errors were corrected by the adjustment of genotype calls or by elimination of genotypes from the data set after re-inspection of the intensity plots. Merlin was used for the detection and removal of unlikely genotype combinations that appeared to have arisen from excessive numbers of recombinations.[71]

Multipoint nonparametric linkage analyses were performed on the genotype data with the use of Merlin,[71] which estimates identical-by-descent allele sharing among affected relatives. To anticipate potential locus heterogeneity within and between populations of different ancestry and potential allele frequency differences, we performed separate linkage analyses on the combined group of European, Ashkenazi Jewish, and Indian ancestry families (Eu/AJ/I) and the East Asian ancestry (E Asian) families, as well as the European ancestry (Eu) families alone. Because parental genotype data were lacking in some of our pedigrees, we used Merlin to form clusters[83] of correlated markers that exhibited pairwise r^2 values greater than 0.16, to ensure that marker-marker linkage disequilibrium was not inflating our multipoint linkage scores.[84] HapMap marker allele frequencies were used for these analyses, though similar results were obtained when allele frequencies were estimated from the founders in our families. Multipoint Kong and Cox exponential nonparametric LOD scores[77] obtained with Whittemore and Halpern's S_{ALL} statistic[76] were then calculated for each marker or marker cluster.

We empirically estimated p-values for our LOD scores by conducting 10,000 gene-dropping simulations under the null hypothesis of no linkage in Merlin[71] and by retaining the LOD scores from the highest independent (separated by ≥ 40 cM) linkage peaks on the autosomes in each replicate. On average, there were about 78 independent linkage peaks per genome scan, resulting in a list of approximately 780,000 LOD score peaks per set of simulations. These simulations used the same marker spacing, clustering, family structures, and informativeness of our study, and we conducted separate sets of simulations on the three subpopulations. The 500th highest LOD score from these simulations was taken to be the empirical threshold for statistical significance (expected to occur in one of every 20 genome scans by chance), and the 10,000th highest LOD score was the empirical threshold for suggestive linkage (expected to occur once in every genome scan by chance).

Linkage analyses

By conducting linkage analysis on the combined set of Eu/AJ/I families, we found that peak LOD scores for two regions of the genome exceeded our empirical threshold for suggestive linkage (LOD = 1.874): chromosome 8q24.21 at rs3057, with a LOD score of 2.330, and chromosome 8q21.11 at rs1007750, with a LOD score of 2.069 (Figure 12A and Table 12). These regions are shown in more detail in Figure 13, and the contributions of individual families to peak LOD scores is detailed in Table 10. In addition, regions on chromosomes 2, 6, 7, 8, 9, 11, and 14 achieved nominal LOD scores greater than 1.0 but did not meet the criteria for suggestive linkage (Table 13).

Examining data for the Eu families alone, we detected one region, with a maximum LOD score at rs3057 on chromosome 8q24.21 (Figure 12B), that showed strong evidence for linkage, having a nonparametric multipoint exponential LOD score of 3.464 (empirical genome-wide p = .0300). This value exceeded the empirical threshold for significant linkage (LOD = 3.231) obtained from 10,000 autosomal gene-dropping simulations. We then used the method of Camp and Farnham[85] to correct for multiple testing. A linear regression of the Eu/AJ/I nonparametric LOD scores versus the corresponding Eu nonparametric LOD scores had an r^2 value of 0.7874, indicating that these two analyses represented 1.213 independent tests. After this correction, the 8q24.21 linkage peak remained significant, with a p value of 0.0364. Three additional regions, on chromosomes 8q21.11 (LOD = 2.236 at rs1007750), 7q22.3 (LOD = 2.074 at rs2028030), and 9p21.3 (LOD = 2.048 at rs2169325), exceeded the empirical threshold for suggestive linkage (LOD = 1.869). Figure 13 shows these regions in more detail. In addition to these significant and suggestive hits (Table 12), several other regions of the genome on chromosomes 2, 4, 6, 10, 11, and 15 had peak LOD scores greater than 1.0 (Table 13). We did not conduct a similar analysis on the eight Ashkenazi Jewish families, because we felt that the sample size was too small to allow linkage detection with any certainty, but it appears that the Ashkenazi Jewish families do not show linkage to the top Eu linkage regions (Table 10). Overall, this linkage analysis indicates that there is a genetic basis for AP in the Eu population.

Using the E Asian families, we observed that no linkage peak exceeded the empirical threshold for suggestive linkage (LOD = 1.822), but regions on chromosomes 1, 3, 7, 13, 18 and 19 had linkage peaks with LOD scores greater than 1.0 (Figure 12C and Table 13). Notably, there was no evidence in the E Asian population for linkage in the region of significant linkage (8q24.21) from the Eu sample set. In fact, the chromosome 7 region was the only E Asian region with a LOD score over 1.0 that showed overlap with linkage peaks observed in the Eu data set (Figure 13A).

In the UCSC genome browser, it appears that four genes lie closest to the top linkage peak on chromosome 8q24.21 in the Eu subset: *GSDMC* (*gasdermin C*), *FAM49B* (a hypothetical protein-coding gene), *ASAP1* (*ArfGAP with SH3 domain, ankyrin repeat and PH domain 1*), and *ADCY8* (*adenylate cyclase 8 (brain)*). *ASAP1* is expressed in a variety of tissues, including the brain,[86] and *ADCY8* is expressed almost exclusively in the brain[87] and is thought to play a role in learning and memory.[88,89] Given that the linkage peak is observed in a single, although broad, population, linkage disequilibrium analysis may help to narrow the interval in the search for genetic variants that lead to AP.

This study bears extension by further recruitment within our own laboratory and replication by other groups interested in this question. Our LOD scores were modest in comparison to the theoretical maximum nonparametric multipoint exponential LOD scores predicted for our samples (maximum exponential LOD scores were 40.03, 34.31, and 12.34 for the Eu/AJ/I, Eu, and E Asian analyses, respectively), and our study was probably underpowered, especially in the case of the E Asian and AJ families. Theoretically, a study of 100 affected sibling pairs could have greater than 90% power to detect linkage, assuming that the $\lambda_s = 10$, the recombination fraction between the marker and trait locus (θ) < .05, and the markers are fully informative, and a similar study of only 40 affected sibling pairs would have 20%–70% power, depending on θ.[58] Though the SNPs in our study were closely spaced ($\theta < 0.0027$ on average), they were not completely informative (average polymorphism information content [PIC] = 0.35).[90] Moreover, we were unable to acquire DNA from informative relatives, such as parents, in some families, so the probabilities that AP relatives share alleles identically by descent were difficult to determine with certainty for these families, thus reducing power further. Despite these considerations, our study was able to detect significant linkage at one locus in the Eu subset, though it was probably underpowered to detect loci that make smaller contributions to predispose individuals to develop AP.

Locus heterogeneity

Because AP is a complex trait and many loci could potentially be involved in its genesis, we also used locus-counting methods[91,92] to evaluate the significance of our linkage results. Again, as with our linkage analyses, we considered the two main sample sets (Eu/AJ/I and E Asian) and the Eu subset separately. First, the top observed linkage regions with LOD scores > 1.0 were arranged in order by rank (r). For each of these observed LOD scores (Z), the number of independent linkage regions (separated by a genetic distance of at least 40 cM) that had LOD scores at least as large as Z in 10,000 autosomal gene-dropping simulations were tallied and divided by 10,000 to determine the average number of times that a LOD score of Z's magnitude was seen in a simulation scan. The final step was to determine the proportion of simulations that had at least as many independent linkage peaks at or above Z as we observed in our linkage analysis (r). If 5% or more of the 10,000 simulations did not contain at least as many independent linkage peaks as our linkage scan did at a LOD score threshold of Z, the excess of linkage peaks at that threshold was considered significant.

Table 13 summarizes the results of this locus-counting analysis for each sample set. In the Eu subset, we observed four independent linkage peaks at or above a LOD score of 2.048; however, on the basis of 10,000 simulations, only 0.68 independent linkage peaks would be expected under the null hypothesis of no linkage at that threshold. The difference between the observed number of linkage peaks and the number expected under the null hypothesis of no linkage based on the simulations was significant ($p = 0.0042$). Similarly, a significant ($p < 0.05$) excess of linkage peaks was observed for the 1st-, 3rd-, 5th-, 10th-, and 11th-ranked independent linkage regions at LOD score thresholds of 3.464, 2.074, 1.723, 1.1, and 1.082, respectively (Table 13). These results indicate that the genetic basis for AP exhibits locus heterogeneity, at least in the Eu population. Though an excess of linkage peaks was also observed when the Eu/AJ/I sample set was used, this difference was not significant. No obvious excess of linkage peaks was found in the E Asian linkage scan with this analysis.

Together, the findings discussed above provide strong evidence that at least one genetic variant promotes the genesis of AP in individuals of European ancestry and that AP probably results from multiple genetic factors that vary both within and between different populations, conclusions that are supported by evidence for linkage in more regions than expected by chance.

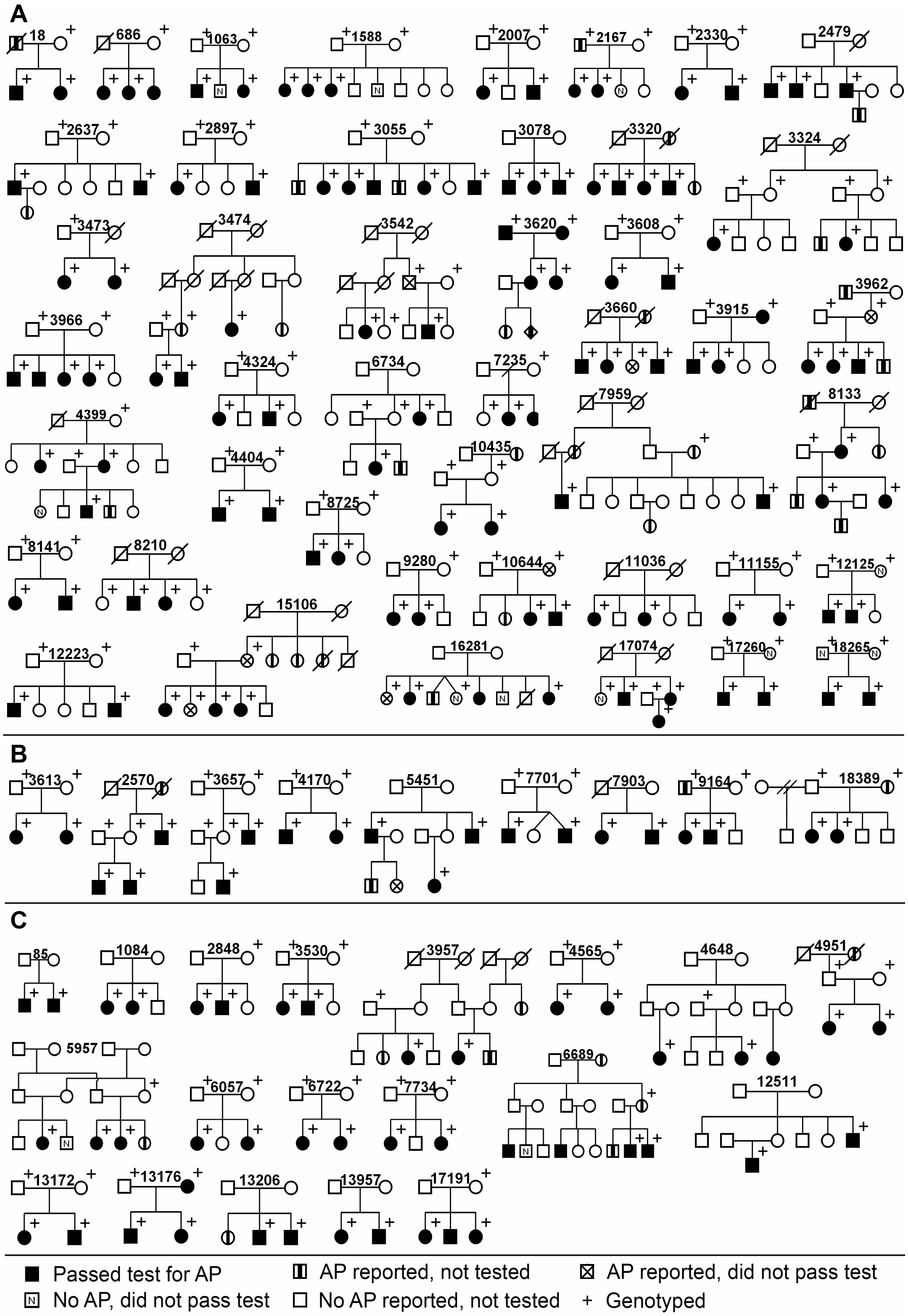


Figure 10. Pedigrees of families used in linkage study. (A) Collection of 45 families of European ancestry. (B) One family of Indian ancestry (3613) and 8 families of Ashkenazi Jewish ancestry. (C) Nineteen families of East Asian ancestry. The key to the symbols is shown at the bottom of the figure.

Table 10. Ethnicities of families and contributions of each family to top linkage regions

Family	Self-Reported Ethnicity	Population[a]	7q22.3[b]	8q21.11[b]	8q24.21[b]	9p21.3[b]
18	British/German	Eu	+	0	-	+
686	German/Italian	Eu	0	+	0	+
1063	German/Serbian	Eu	+	-	+	+
1588	Finnish	Eu	-	+	+	+
2007	Polish/Austrian/Swedish/Scottish	Eu	0	+	+	+
2167	German/Danish/English/Irish	Eu	+	+	0	0
2330	Ukrainian/Russian	Eu	0	0	0	-
2479	French/Norwegian/German	Eu	+	+	+	+
2637	German/Swiss	Eu	0	-	+	-
2897	Polish/Scottish/French Canadian	Eu	-	+	0	+
3055	Swedish	Eu	+	+	+	+
3078	Caucasian	Eu	+	+	-	+
3320	English	Eu	-	-	+	+
3324	Dutch	Eu	+	+	-	+
3473	Caucasian	Eu	0	-	-	0
3474	German	Eu	+	-	-	+
3542	Polish	Eu	+	-	+	-
3608	Mennonite (German/Dutch)	Eu	0	0	-	+
3620	French/German (¼ AJ)	Eu	+	-	0	+
3660	English/Slovakian/German/Austrian	Eu	+	+	+	-
3915	¾ Scottish/English/German, ¼ Japanese	Eu	+	+	+	-
3962	French Canadian	Eu	-	+	+	+
3966	European	Eu	+	+	-	+
4324	Swedish/Norwegian/Polish/Danish	Eu	+	-	+	0
4399	Caucasian	Eu	+	+	+	+
4404	Caucasian	Eu	0	+	0	+
6734	Italian/French/English/Irish/Scottish	Eu	+	+	+	-
7235	Italian/Slovak/French/English/Dutch	Eu	+	+	0	-
7959	Caucasian	Eu	+	+	+	+
8133	Scottish/German/Native American	Eu	-	+	+	+
8141	Swedish/German/Bulgarian	Eu	0	0	-	+
8210	English/Irish/Scandinavian	Eu	0	0	+	-
8725	English/Scottish	Eu	+	-	+	+
9280	German/Italian	Eu	0	+	-	0
10435	Irish/French/Scottish/German	Eu	-	0	+	0
10644	Dutch	Eu	+	0	+	0
11036	German	Eu	-	0	+	-
11155	Italian/German/Irish	Eu	0	+	+	0
12125	Hungarian/German/Irish	Eu	-	+	+	-
12223	Caucasian LDS (Mormon)	Eu	+	+	+	0
15106	English	Eu	-	+	+	-
16281	English/Irish	Eu	0	+	-	0
17074	English/Irish/Dutch/Scottish	Eu	+	-	-	-
17260	British/Swedish	Eu	0	0	+	+
18265	British/German/Italian	Eu	0	+	+	0
3613	Indian	I	0	+	0	+
2570	Ashkenazi Jewish	AJ	+	-	-	+
3657	Ashkenazi Jewish	AJ	-	-	-	-
4170	AJ (Russian)	AJ	-	+	0	0
5451	AJ (Russian)	AJ	-	+	-	+
7701	AJ (Hungarian/Russian/Lithuanian)	AJ	0	0	0	0
7903	Ashkenazi Jewish	AJ	0	0	0	0
9164	Jewish (Middle Eastern/Romanian)	AJ	0	0	0	0
18389	Ashkenazi Jewish	AJ	0	0	0	-

Family	Ethnicity	Population	7q22.3[b]	8q21.11[b]	8q24.21[b]	9p21.3[b]
85	Chinese	E Asian	+	0	0	-
1084	Chinese	E Asian	-	-	0	-
2848	Chinese	E Asian	-	+	0	+
3530	Chinese	E Asian	+	+	+	-
3957	Japanese	E Asian	+	+	-	+
4565	Chinese	E Asian	-	0	-	0
4648	Chinese	E Asian	-	+	0	-
4951	Filipino	E Asian	+	0	0	-
5957	Taiwanese	E Asian	+	+	0	+
6057	Filipino	E Asian	0	+	0	0
6722	Chinese	E Asian	+	-	+	-
6689	Chinese	E Asian	-	0	+	+
7734	Chinese	E Asian	+	-	-	-
12511	Filipino/Chinese	E Asian	+	+	-	+
13172	Chinese	E Asian	+	+	-	-
13176	Taiwanese	E Asian	+	+	+	+
13206	Chinese	E Asian	+	-	0	0
13957	South Korean	E Asian	-	0	-	+
17191	Taiwanese	E Asian	+	-	+	-

[a] All families of European (Eu), Ashkenazi Jewish (AJ), and Indian (I) ancestry are included in the Eu/AJ/I sample set, while the Eu sample set excludes families of Ashkenazi Jewish and Indian descent.

[b] Contributions were considered positive (+) if they were ≥0.01 multipoint nonparametric exponential logarithm of the odds (LOD), negative (-) if they were ≤-0.01 LOD, and neutral (0) if they were between -0.01 and 0.01 LOD.

Refer to Figure 10 for pedigrees of these families.

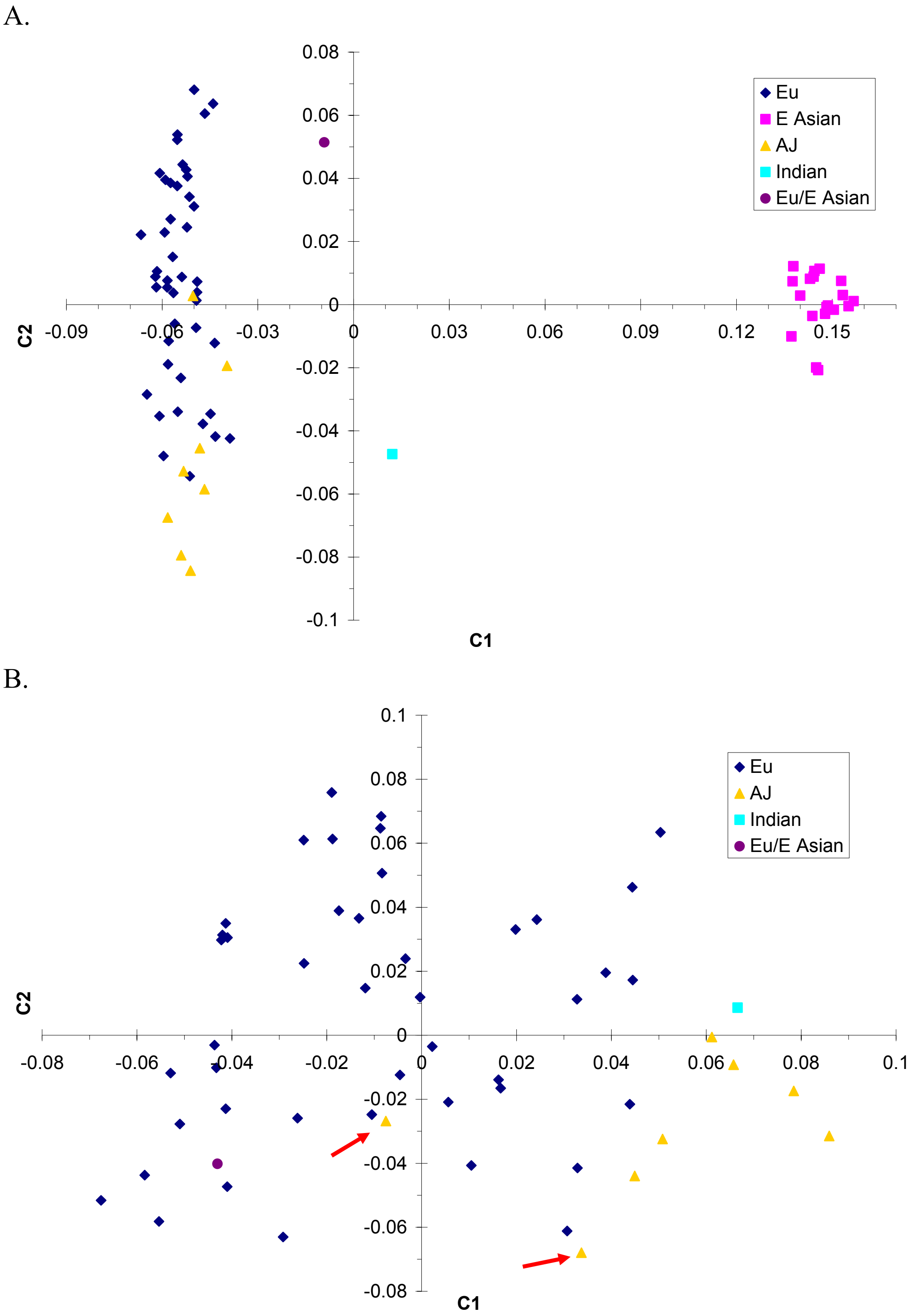


Figure 11. Multidimensional scaling plot of AP family probands based on pairwise identity-by-state (IBS) distances. IBS distances and plot coordinates were generated separately using probands from (A) all families or (B) only Eu, AJ, and I families in Plink.[82] In B, the red arrows point to two probands that likely have less than 100% AJ ancestry and some Eu ancestry based on genetic data obtained for our AJ whole-genome association study (Figure 16).

Table 11. Description of families used in linkage analysis

	Eu/AJ/I	**Eu**[a]	**E Asian**
No. of families	54	45	19
No. of individuals genotyped	220	184	61
No. of AP individuals genotyped	128	108	40
No. of AP sibling pairs	73	65	16
No. of AP avuncular pairs	8	3	1
No. of AP cousin and distant pairs	5	5	4
No. of AP relative pairs[b]	86	73	21

[a] The European descent (Eu) sample set is a subset of the Eu/AJ/I sample set, excluding one Indian and eight Ashkenazi Jewish families.

[b] AP parent-child pairs were not included in the relative-pairs count.

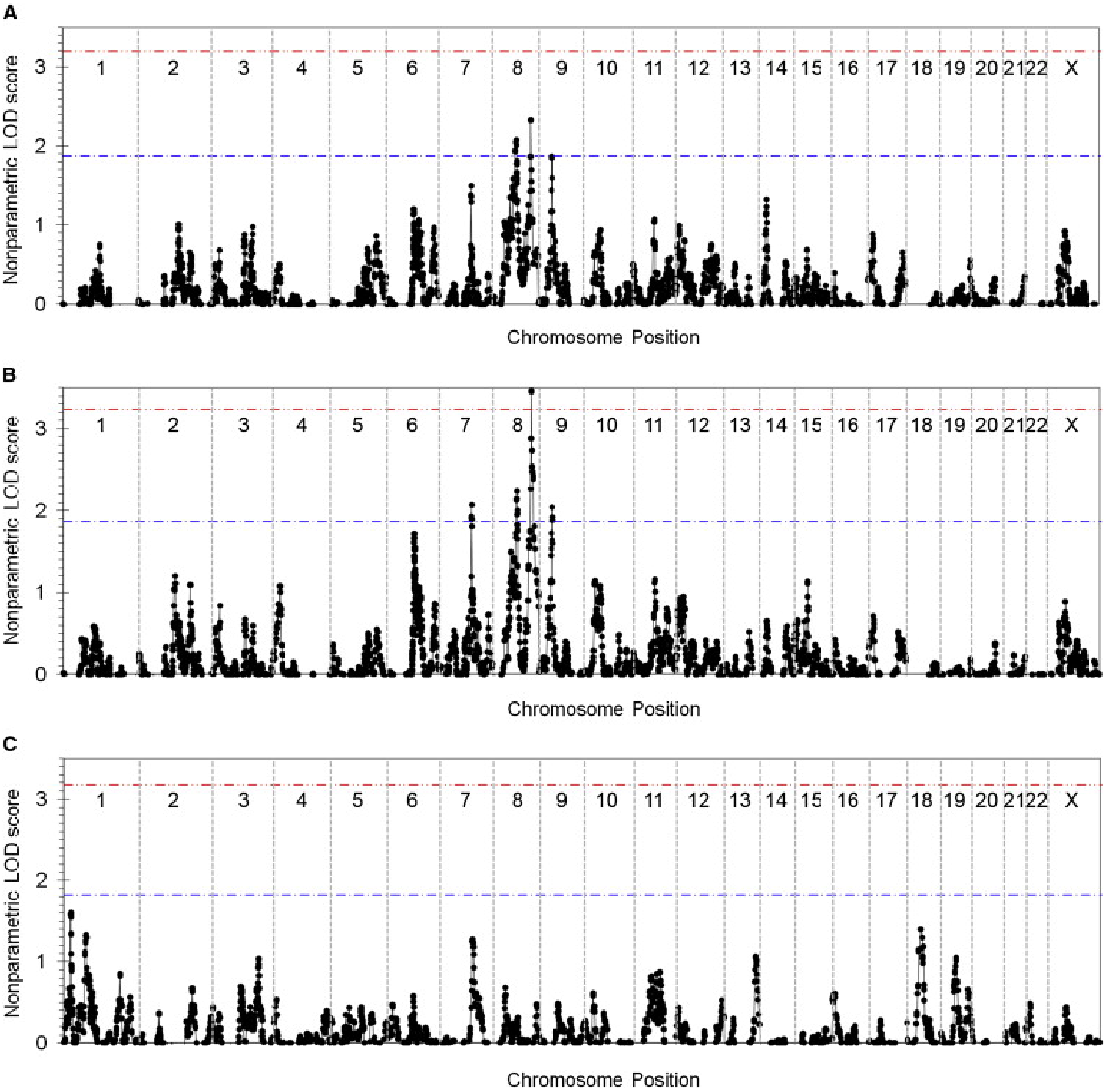


Figure 12. Results of whole-genome linkage analysis. Nonparametric multipoint exponential LOD scores were calculated for every marker or marker cluster position across the genome with the use Merlin for (A) the 54 families of European, Ashkenazi Jewish, and Indian (Eu/AJ/I) descent, (B) a subset of 45 families of mixed European (Eu) descent, and (C) 19 families of East Asian (E Asian) descent. Only nonnegative LOD scores are shown. Red and blue lines indicate empirical thresholds for significant and suggestive linkage, respectively, with 10,000 gene dropping simulations used.

Table 12. Significant and suggestive chromosome regions from multipoint nonparametric linkage analysis.

Sample Set	Region	Marker	deCODE cM	LOD[a]	Emp p Value[b]	Interval Size (Mb)[c]	Flanking Markers
Eu/AJ/I	8q21.11	rs1007750	86.732	2.069	0.6490	22.86	rs997493-rs10105219
Eu/AJ/I	8q24.21	rs3057	139.741	2.330	0.3611	6.01	rs1562435-rs2102861
Eu	7q22.3	rs2028030	117.774	2.074	0.6402	4.04	rs887882-rs1013920
Eu	8q21.11	rs1007750	86.732	2.236	0.4500	11.75	rs695167-rs716349
Eu	8q24.21	rs3057	139.741	3.464	***0.0300***[d]	5.54	rs755520-rs2102861
Eu	9p21.3	rs2169325	46.478	2.048	0.6786	7.91	rs748530-rs9103

[a] The top multipoint nonparametric exponential LOD scores from linkage analysis of the European, Ashkenazi Jewish, and Indian ancestry sample set (Eu/AJ/I) and the subset of families of European ancestry (Eu).

[b] Empirical genome-wide p values were estimated for each sample set independently by calculating the average numbers of independent linkage peaks expected under the null hypothesis of no linkage per genome scan, with 10,000 autosomal simulations run.

[c] Intervals are LOD − 1.

[d] Bold italics denote significant results ($p < 0.05$).

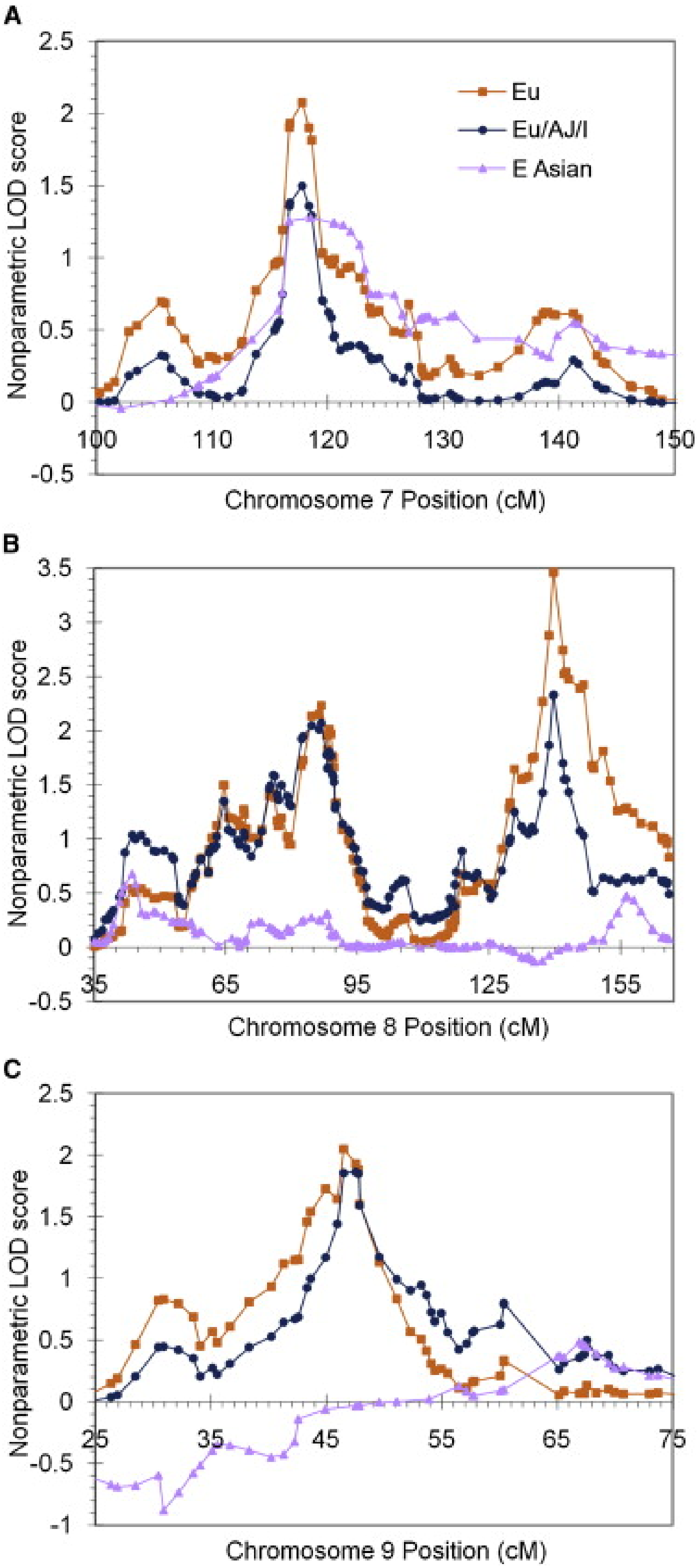


Figure 13. Suggestive and significant linkage regions. Multipoint nonparametric exponential LOD scores at each marker or marker cluster position were calculated with the use of data from 54 families of European, Ashkenazi Jewish, and Indian descent (Eu/AJ/I), the subset of 45 families of European descent (Eu), and 19 East Asian families (E Asian). Genetic distance is measured in deCODE cM. Results are shown for (A) chromosome 7, (B) chromosome 8, and (C) chromosome 9.

Table 13. Evaluation of nonparametric peak LOD scores > 1.0 by locus counting

Sample Set[a]	**Chr**	**deCODE cM**	**Observed LOD (Z)**	**Rank (r)**	**No. of LODs ≥ Z Per Sim Scan**[b]	**Prop of Sims with r LODs ≥ Z**[c]
Eu/AJ/I	8	139.741	2.330	1	0.3611	0.3034
Eu/AJ/I	8	86.732	2.069	2	0.6490	0.1372
Eu/AJ/I	9	47.565	1.864	3	1.0240	0.0878
Eu/AJ/I	7	117.774	1.499	4	2.2296	0.1852
Eu/AJ/I	14	27.385	1.322	5	3.2538	0.2299
Eu/AJ/I	6	99.679	1.201	6	4.2168	0.2462
Eu/AJ/I	11	78.450	1.080	7	5.4576	0.3032
Eu/AJ/I	8	43.577	1.038	8	5.9693	0.2498
Eu/AJ/I	2	145.803	1.007	9	6.3761	0.1898
Eu	8	139.741	3.464	1	***0.0300***[d]	***0.0293***[d]
Eu	8	86.732	2.236	2	0.4500	0.0750
Eu	7	117.774	2.074	3	0.6402	***0.0262***[d]
Eu	9	46.478	2.048	4	0.6786	***0.0042***[d]
Eu	6	99.679	1.723	5	1.3904	***0.0128***[d]
Eu	2	130.661	1.205	6	4.2078	0.2494
Eu	11	78.450	1.160	7	4.6294	0.1839
Eu	10	41.725	1.147	8	4.7538	0.1072
Eu	15	47.975	1.133	9	4.8958	0.0590
Eu	2	185.422	1.100	10	5.2524	***0.0378***[d]
Eu	4	25.557	1.082	11	5.4560	***0.0209***[d]
E Asian	1	25.394	1.606	1	1.6028	0.7996
E Asian	18	49.984	1.399	2	2.5007	0.7123
E Asian	1	81.597	1.326	3	2.9287	0.5621
E Asian	7	118.517	1.277	4	3.2562	0.4078
E Asian	13	114.683	1.064	5	5.1098	0.5854
E Asian	19	55.795	1.051	6	5.2597	0.4296
E Asian	3	168.852	1.040	7	5.3825	0.2912

Both [b] and [c] were calculated for all observed independent linkage peaks with LOD scores exceeding 1.0.

[a] Simulations were conducted independently for the European, Ashkenazi Jewish, and Indian sample set (Eu/AJ/I), for the European ancestry subset (Eu), and for the East Asian (E Asian) sample set.

[b] Average numbers of independent linkage peaks per genome scan observed under the null hypothesis of no linkage in 10,000 autosomal simulations.

[c] The proportion of 10,000 autosomal simulations that had at least *r* linkage regions with LOD scores greater than or equal to the observed LOD score.

[d] Bold italics denote significant results ($p < 0.05$).

V. COMBINED LINKAGE STUDY USING SNP AND MICROSATELLITE MARKERS

Since we had collected genotype data for many of our families at both microsatellite and SNP markers, we analyzed all data in a combined linkage analysis. The pedigrees that were genotyped with microsatellite and SNP markers are shown in Figures 3, 5, and 10. We first conducted multipoint nonparametric exponential linkage analyses in Merlin,[71] using the markers' actual and interpolated[73] genetic distances on the deCODE map.[57] We did not cluster any markers in these analyses, in contrast to our analyses of the SNP markers alone (Chapter IV). Separate analyses were conducted for families of European ancestry (Eu), families of East Asian ancestry (EAsian), and families of European ancestry, Ashkenazi Jewish ancestry, or Indian ancestry (Eu/AJ/I). The results of these analyses (Figure 14) closely resembled those when using SNP markers alone (Figure 12), though the magnitudes of some of the peak LOD scores were slightly different (Tables 13 and 14). The results from the Eu and Eu/AJ/I linkage analyses were quite similar, but the only linkage region with a LOD score greater than 1.0 in Eu, Eu/AJ/I, and EAsian families was the one on chromosome 7 (Table 14).

We also used the combined microsatellite and SNP genotype data to conduct multipoint parametric linkage analyses using previously described models (Table 8) on the Eu and EAsian families (Figure 15) similar to those described for microsatellite markers in Chapter III and depicted in Figure 8. By comparing the magnitudes of the heterogeneity LOD scores from different models, we hoped to get some insight into the pattern of inheritance of the candidate AP-predisposing genomic regions. The best-fitting models are summarized in Table 14. As with our earlier microsatellite linkage analysis findings, the Eu chromosome 8 region near 140 cM fit best with a recessive model, and the Eu chromosome 9 region fit best with a dominant model.

Overall, it was encouraging to see that our top candidate linkage regions from the SNP linkage study did not disappear upon addition of the microsatellite marker data. Those potential linkage regions that seemed promising from the microsatellite linkage study but did not replicate upon addition of SNP markers, such as the region on chromosome 4p, could potentially have been artifacts due to the poor quality of the microsatellite marker data. Unfortunately, we could not conclusively determine why it wasn't substantiated because much of the raw data from the microsatellite study was no longer available.

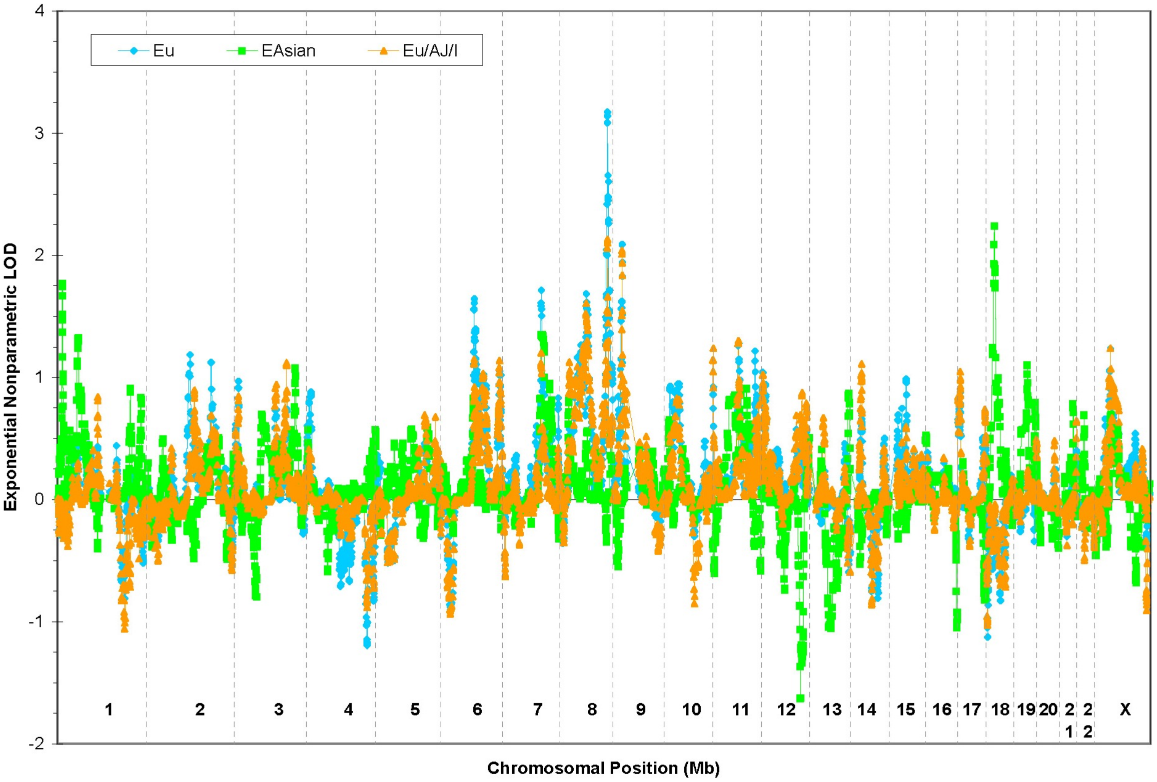


Figure 14. Nonparametric multipoint exponential linkage analyses of families of European ancestry (Eu), East Asian ancestry (EAsian), and European, Ashkenazi Jewish, or Indian ancestry (Eu/AJ/I) using combined microsatellite and SNP genotype data.

Table 14. Summary of most promising linkage regions from combined SNP and microsatellite linkage analysis. Genetic distances are deCODE cM. LOD scores are nonparametric exponential.

Pop	Chr	cM	Marker	LOD	Best Model	Alt Model
Eu	2	130.661	rs1880542	1.187	ComMix	RareDom
Eu	2	185.422	rs2007326	1.125	RareDom	ComMix
Eu	6	99.665	rs1979797	1.645	ComRec	ComMix
Eu	6	173.291	rs1954948	1.054	ComMix	ComRec
Eu	7	117.774	rs2028030	1.714	ComMix	RareDom
Eu	8	84.489	rs6988179	1.686	ComMix	RareDom
Eu	8	139.790	D8S284	3.173	ComRec	ComMix
Eu	9	46.478	rs2169325	2.091	ComDom	RareDom
Eu	11	78.937	rs593753	1.281	RareDom	ComDom
Eu	11	120.300	rs947889	1.218	ComMix	ComRec
Eu	12	8.430	D12S372	1.043	ComMix	ComRec
Eu	X	68.268	rs1451512	1.236	RareDom	ComMix
Eu/AJ/I	3	151.039	rs765695	1.121	-	-
Eu/AJ/I	6	99.665	rs1979797	1.150	-	-
Eu/AJ/I	6	173.291	rs1954948	1.138	-	-
Eu/AJ/I	7	117.774	rs2028030	1.200	-	-
Eu/AJ/I	8	45.694	rs388047	1.126	-	-
Eu/AJ/I	8	84.489	rs6988179	1.608	-	-
Eu/AJ/I	8	139.790	D8S284	2.130	-	-
Eu/AJ/I	9	46.695	D9S1121	2.040	-	-
Eu/AJ/I	11	1.259	D11S4046	1.241	-	-
Eu/AJ/I	11	78.750	D11S1314	1.302	-	-
Eu/AJ/I	12	14.723	rs248881	1.024	-	-
Eu/AJ/I	14	27.385	rs2273171	1.112	-	-
Eu/AJ/I	17	18.837	rs1848550	1.047	-	-
Eu/AJ/I	X	68.268	rs1451512	1.239	-	-
EAsian	1	24.728	rs761162	1.767	RareDom	ComDom
EAsian	1	81.597	rs927612	1.324	RareDom	ComMix
EAsian	3	168.852	rs11921535	1.075	ComRec	ComDom
EAsian	7	116.747	rs257376	1.351	ComRec	ComDom
EAsian	18	49.762	rs1185007	2.237	ComRec	ComDom
EAsian	19	55.795	rs977708	1.104	ComMix	ComDom

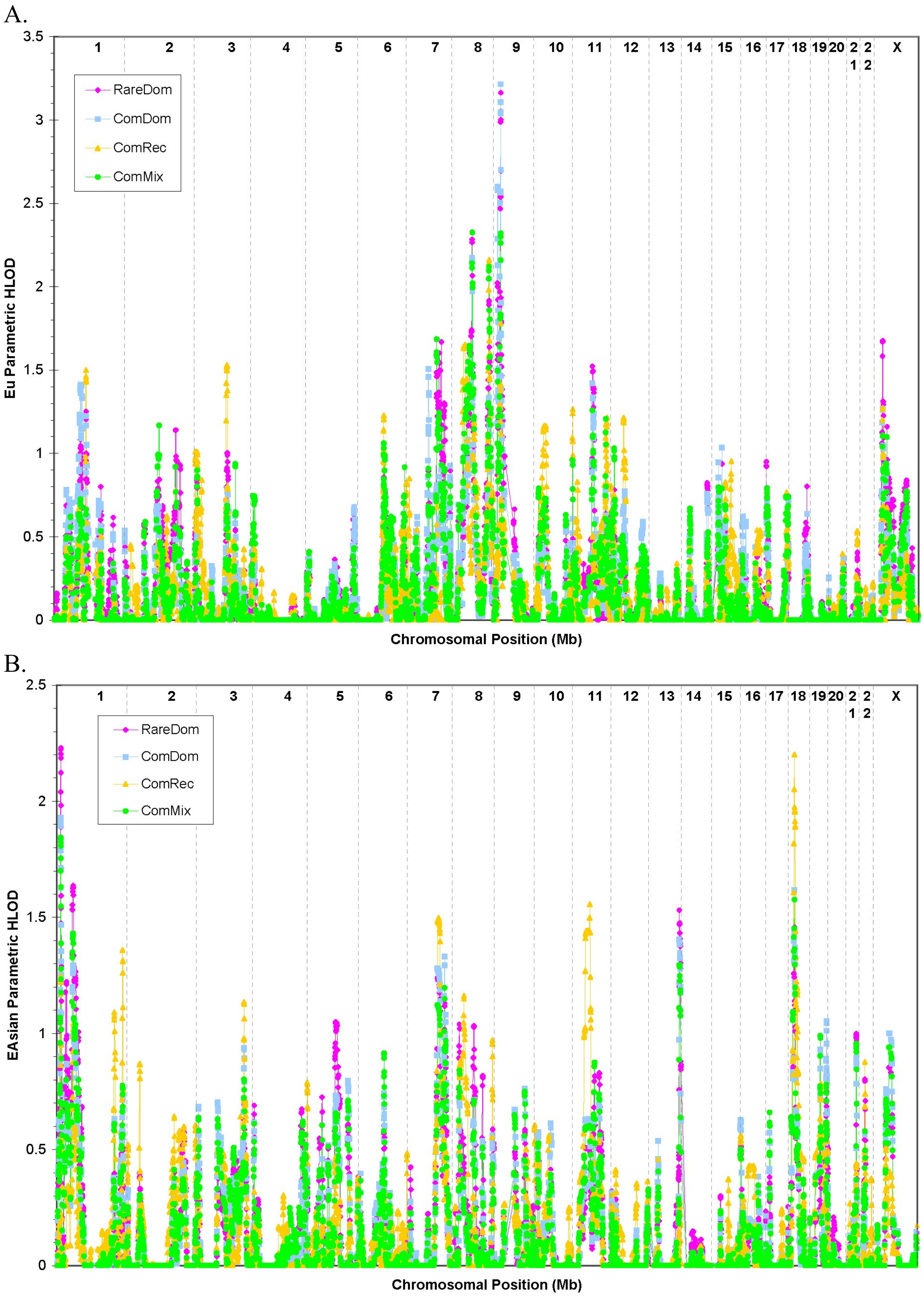


Figure 15. Parametric linkage analyses of (A) Eu and (B) EAsian families using SNP and microsatellite marker data.

VI. GENOME-WIDE ASSOCIATION STUDY IN ASHKENAZI JEWS

To complement the linkage approaches used in the previous chapters, we employed a genome-wide association study (GWAS) to detect additional AP-predisposing genetic variants. We chose to try a GWAS approach because we could theoretically detect common variants of smaller effect size than we could in a linkage study. Furthermore, recruitment for a case-control study could be more fruitful because we could use unrelated AP individuals, not just AP individuals that have family members with AP.

We chose to conduct our GWAS in a special population, the Ashkenazi Jews (AJ), because though that population is closely related to others of European ancestry, it also has unique characteristics that could make a GWAS particularly effective. The AJ are a recently expanded founder population[93] that is relatively genetically homogeneous, making it better suited for genome-wide association studies of complex traits than outbred populations, because a smaller number of genetic variants may be contributing to the trait in the AJ as compared to outbred populations.[94] The Ashkenazi Jews are an especially attractive AP study population because of their strong musical tradition and their sizeable population the U.S. In addition, stretches of LD are longer on average in the Ashkenazi Jewish population as compared to outbred populations,[95] making regions inherited identically by descent easier to detect. Thus, fewer markers and smaller sample sizes would theoretically be necessary to successfully locate shared haplotypes in the Ashkenazi Jewish population than in an outbred population. Any interesting variants that would be detected in this isolated population could then be investigated to determine if they are involved in the development of absolute pitch in a wider range of ethnic backgrounds.

Participants and Genotyping

Ashkenazi Jewish participants were identified from the large pool of individuals who had entered our study over the past 10 years and had been surveyed and tested online or via a paper-based survey and CD-based test. Participants who tested with AP were asked via e-mail to describe the ethnicity and country of ancestry of each of their four grandparents, if known. This self-report information allowed us to select AP individuals of alleged Ashkenazi Jewish ancestry, but in some cases we were not sure whether these individuals were 100% AJ or less than 100% AJ until we genotyped them with a dense array of markers (Figure 16). With the introduction of the most recent version of our online survey in February 2008, we asked every participant, regardless of AP status, about their ethnic ancestry. This allowed us to find control AJ participants who had musical training before the age of 7 but did not develop absolute pitch. DNA samples were then collected from AJ participants via mouthwash, saliva, or blood samples.

Our DNA samples were genotyped by the UCSF Genomics Core Facility on Illumina Infinium HumanHap 550K-Duo version 3 or 610K-Quad BeadChips. A total of eight participants with AP were genotyped on the 550K chips, while 35 participants with AP and 13 participants without AP were genotyped on the 610K chips.

In addition to our small number of control participants, we sought data from individuals of Ashkenazi Jewish descent who had been genotyped in other studies. Though we did not have AP survey or test data from these additional participants, we assumed the majority of them did not have AP because AP is rare in the general population as well as the AJ population. The first set of data from external control participants was taken from the collection of normal individuals enrolled in the New York Health Project.[96] Illumina Infinium HumanHap 300K BeadChip genotype data for 392 individuals were downloaded from the InTraGenDB population genetics database (https://intragen.c2b2.columbia.edu/). These individuals were all listed as having Ashkenazi Jewish ancestry in the database. In addition, we acquired genotype data from 48 individuals, each with four Ashkenazi Jewish grandparents, from Anna Need and David Goldstein at Duke University.[97] Eight of these individuals were genotyped on HumanHap 550K version 1 BeadChips, three were genotyped on 550K-Duo version 3 chips, and the remaining 37 were genotyped on 610K-Quad chips.

Aside from the New York Heath Project data, in which the SNP genotypes were already called, the first step was to import the raw intensity data into Illumina's BeadStudio software and call the genotypes for each SNP based on chip intensity data. Intensity data from five of our participants with AP were of low quality and were discarded from subsequent analyses. Called genotype data from all of the remaining participants was then imported into Plink for subsequent analyses.[82]

Participant Pruning

In order to reduce population stratification, we needed to determine if the cases and controls in our study were genetically different from individuals who were known to have four Ashkenazi Jewish grandparents, so we estimated genetic distances between individuals in our study. To accomplish this, we needed to refine the SNP list to use when calculating the genetic distances. We first excluded SNPs with minor allele frequencies less than 1% and SNPs that were genotyped in less than 90% of samples.

This left 291,485 autosomal SNPs. To eliminate large clusters of SNPs in linkage disequilibrium that could disproportionately contribute to our analysis, we then removed SNPs that had an $r^2 > 0.3$ with at least one other SNP in a 1,500 SNP window (with a step size of 150 SNPs)[97] using the indep-pairwise command in Plink.[82]

The genotype data from the remaining 115,455 SNPs were used to create a matrix of pairwise identity by state (IBS) differences among all of the individuals of interest. To determine if samples of reported 100% Ashkenazi Jewish ancestry had substantial genetic contributions from non-Jewish individuals of European ancestry, publicly available Illumina Infinium HumanHap 610K-Quad genotype data for 73 HapMap CEPH/CEU (Utah residents with northern and western European ancestry) individuals were downloaded (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE17205) and included in this analysis.

The IBS distances calculated above were used to generate multi-dimensional scaling (MDS) plots. The first attempt to create plots revealed two closely related pairs of individuals (D0009273 & D0000784; D0005827 & D0004077) in the New York Health project dataset (data not shown). After D0000784 and D0005827 were removed from the analysis, the MDS plots revealed a large cluster of individuals of apparent 100% AJ ancestry and also some individuals with mixed ancestry (Figure 16). We calculated the mean and standard deviation of the first dimension coordinates for samples that were known to have four Ashkenazi Jewish grandparents[97] and excluded 61 non-CEU samples that were at least three standard deviations from that mean, leaving 428 individuals suspected to have full AJ ancestry for association analyses.

SNP Pruning

Since the GWAS included genotype data generated from four different Illumina platforms and three different laboratories, we also needed to exclude SNPs that had inconsistent calls between platforms or laboratories, so we conducted some control allelic association analyses using samples from individuals that were not known to have absolute pitch.

Our first analysis involved 61 individuals who were not known to have AP and were genotyped on the 550K, 550K-Duo, or 610K-Quad platforms as "cases" and 332 individuals with unknown AP who were genotyped on the 300K platform as "controls". The QQ-plot generated in R[98] using -log(p) values generated in Plink[82] revealed that a subset of markers have different allele frequencies in controls genotyped on the 550K+ platforms versus the 300K platform (Figure 17A). Similarly, some markers had different allele frequencies in 11 individuals genotyped on the 550K and 550K-Duo platforms versus 50 individuals genotyped on the 610K-Quad platform (Figure 17B). In contrast, no markers showed dramatic allele frequency differences between the 13 individuals genotyped at UCSF on the 610K-Quad platform and the 37 individuals genotyped at Duke on the 610K-Quad platform (Figure 17C).

We initially thought that these analyses would be sufficient to eliminate platform or genotyping center-biased markers from our dataset. However, after an initial allelic association analysis using AP individuals as cases and non-AP or unknown AP individuals as controls (data not shown), we noticed that the allele frequencies of some of our top hits in non-AP possessors differed on the 610K-Quad platform depending on the barcode of the chip the sample was processed on. Specifically, some marker allele frequencies differed between 15 individuals with no or unknown AP that were genotyped on 610K-Quad chips with serial numbers greater than or equal to 4732268054 compared to 35 individuals with no or unknown AP that were genotyped on 610K-Quad chips with serial numbers less than or equal to 4637092163 (Figure 17D).

To eliminate the most biased markers, we removed the 100 markers from each of the 550K+ versus 300K, 550K versus 610K, and 610K late versus 610K early analyses with the lowest p-values in the control allelic associations. (Due to a little overlap, this totaled 291 markers.) Since the New York Health Project samples were only genotyped with about 300,000 markers while the rest of the samples were genotyped with over 550,000, we did not want to simply eliminate SNPs that were not genotyped in 95% or more of the samples, because we would lose a lot of potentially important information that way. Instead, we removed the markers that were not on the 300K chip only if they were genotyped in less than 95% of the 96 samples genotyped with at least 550K SNPs. We also removed markers that were on the 300K chip if they were genotyped in less than 95% of the complete 428 sample set and/or the 96 samples genotyped with at least 550K SNPs. This removed a total of 113,971 markers. We also removed the 127 markers that appeared to be out of Hardy-Weinberg equilibrium, with p-values less than or equal to 0.00001, and the 5,229 remaining markers that had a minor allele frequency (MAF) less than 1%. After the initial 620,901 markers went through all of these pruning steps, 505,485 markers remained for association analysis.

Allelic Association in AJ

Allelic association analysis of the 35 AP cases and 393 non-AP or unknown AP controls was performed using Plink,[82] and only one marker exceeded the threshold for statistical significance after Bonferroni correction for multiple testing, while two additional markers exhibited suggestive association (Figure 18, Table 15). The observed unadjusted –log(p) values were plotted against the expected values on a quantile-quantile plot in R,[98] and only a few markers had –log(p) values that were moderately greater than those expected by chance (Figure 19). There did not appear to be markers near the most associated markers that also showed evidence for association, so either the most associated alleles were not in linkage disequilibrium with nearby genotyped markers, or they could be false positives. Since our study was very small for a GWAS due to limited time and money, we knew that it was only powered to detect variants of moderate to large effect sizes. Thus, it was not surprising that few promising hits resulted from this analysis.

The marker with significant association to absolute pitch, rs3735251, is located in an intron of *AGR3*. *AGR3* (anterior gradient protein 3) is primarily expressed in the trachea, fetal lung, and colon[99] and may play a role in breast cancer.[100] Though there is nothing known about the gene that suggests it may play a role in brain development and/or function, it is not outside the realm of possibility. The two markers with suggestive association to AP are unlikely to account for much of AP predisposing variation because their allele frequencies were quite low in our AP cases (Table 15). The first, rs2039290, is located in an intron of *SLC1A1*, a glutamate transporter that may play a role in obsessive-compulsive disorder.[101] The second, rs8065590, is located intergenically, approximately 30 Kb away from both *NOL11* and *BPTF*. *NOL11* is a nucleolar protein expressed primarily in white blood cells, while *BPTF* (bromodomain

PHD finger transcription factor) is expressed in a variety of tissues including white blood cells and the brain.[99]

Association in additional participants

The next step was to see if any of our top markers replicated in a separate population. Since we had not recruited enough participants of Ashkenazi Jewish ancestry for a replication study, we decided to see whether our associations replicated in an independent sample of participants of European ancestry. These participants were genotyped for the SNPs on the Sequenom iPlex Gold Genotyping platform by the UCSF Cancer Center Genome Analysis Core. Unfortunately, the five most associated SNPs from the GWAS were not significantly associated with AP in individuals of European ancestry when using 31 AP possessors and 24 musically trained individuals without AP (Table 16). Without replication, it is difficult to determine if the association signals were false or real.

We also investigated whether the top five most associated SNPs showed association with absolute pitch within our few Ashkenazi Jewish families by genotyping all AP individuals in those families using the Sequenom platform. Unfortunately, the results were largely inconclusive due to a lack of parental information and a lack of informativeness of the markers in the majority of the families.

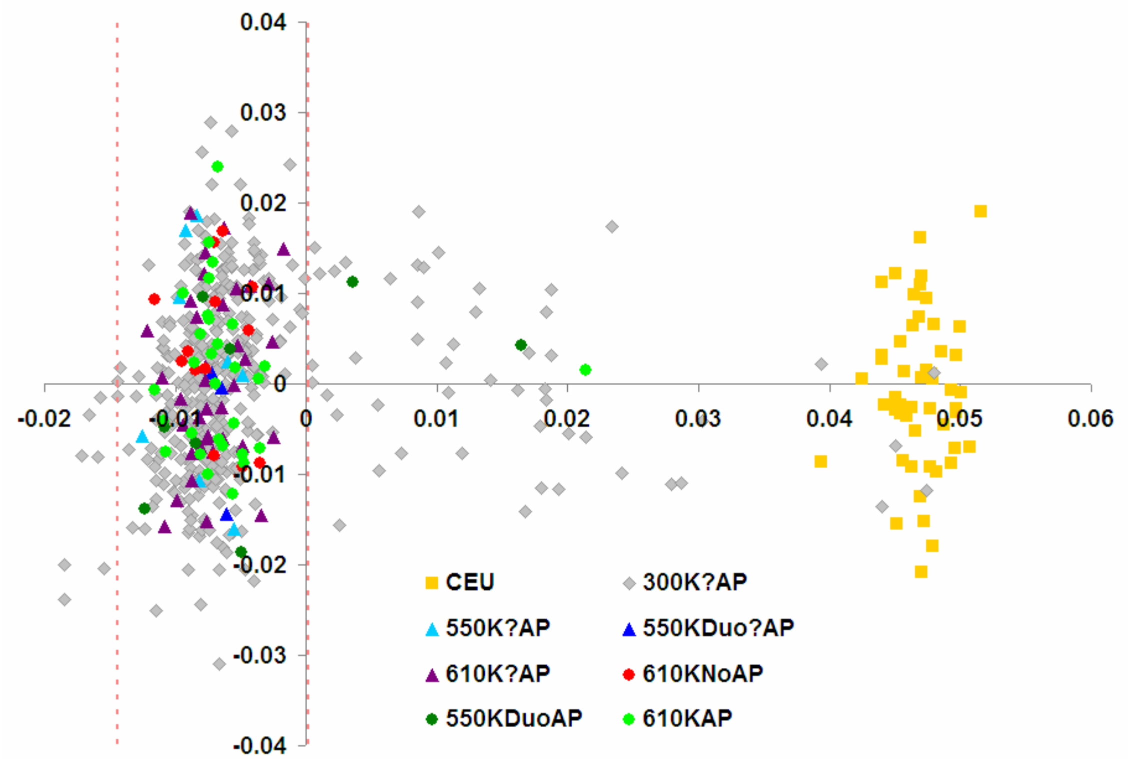


Figure 16. First two dimensions of multidimensional scaling plot based on pairwise IBS distances between individuals, with CEU individuals for reference. Vertical dashed lines indicate +/- 3 SD from the mean of the first dimension coordinates in samples known to have 4 Ashkenazi Jewish grandparents (shown here as triangles).

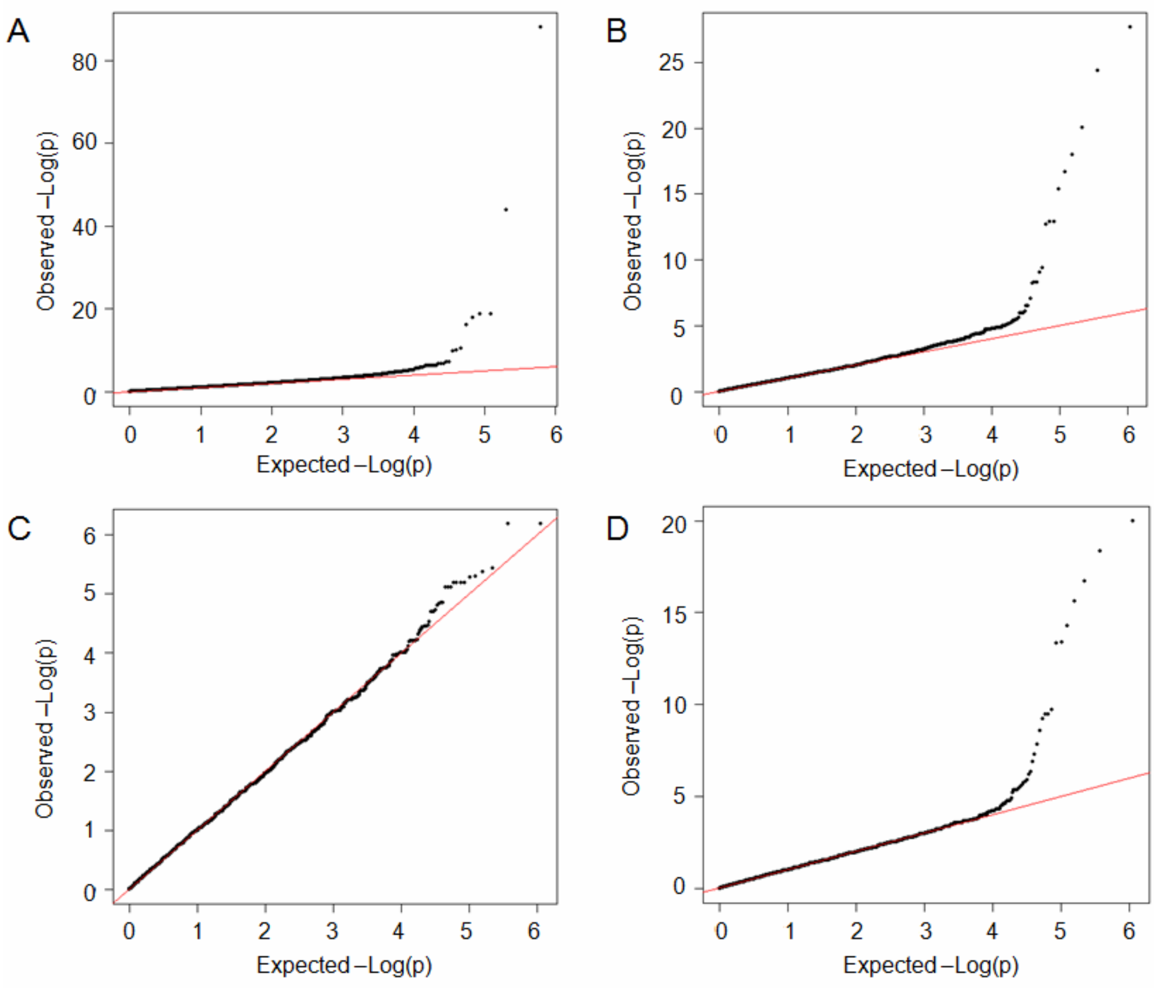


Figure 17. Quantile-quantile plots for control inter-platform, inter-institution, and intra-platform association studies. Allelic association analyses were conducted using (A) 393 control individuals (61 550+K vs. 332 300K) (B) 61 control individuals (11 550K vs. 50 610K) (C) 50 control individuals (13 UCSF 610K vs. 37 Duke 610K) or (D) 50 control individuals (15 late vs. 35 early 610K serial numbers).

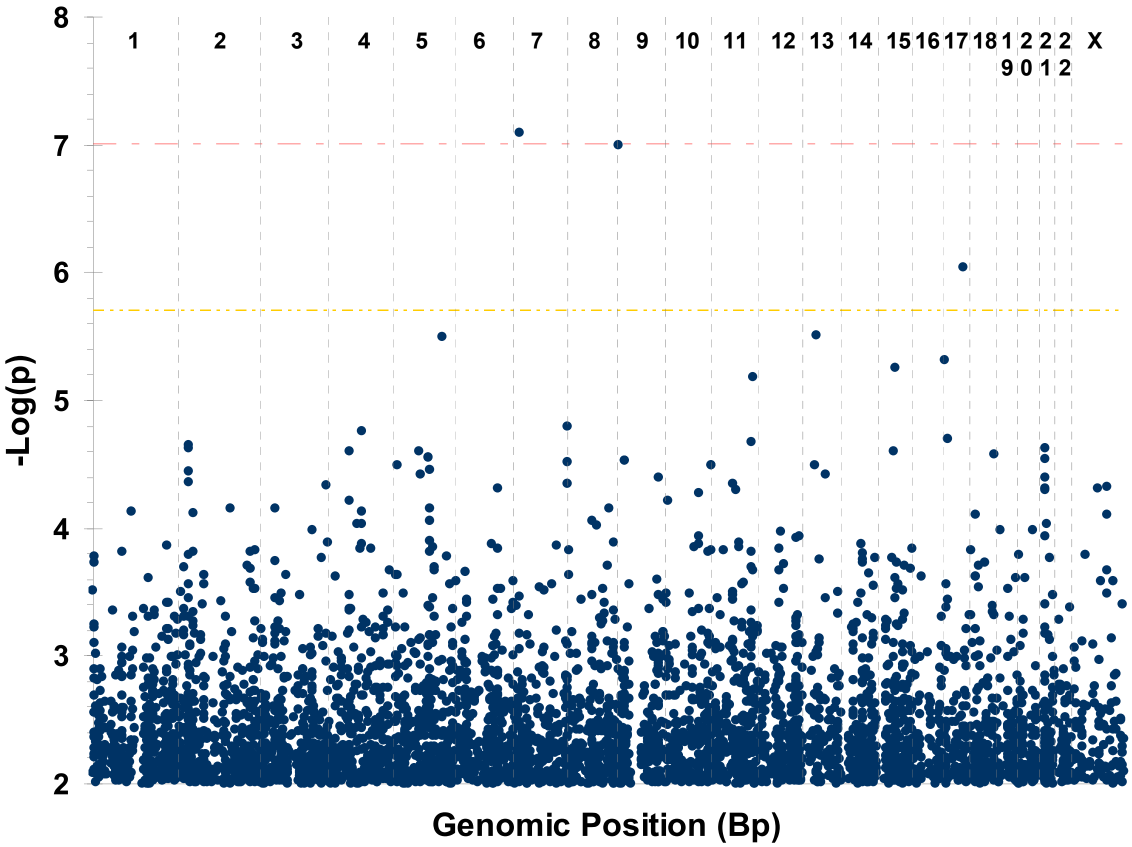


Figure 18. Results of case-control allelic association study. Horizontal red line is the Bonferroni significance threshold after correcting for 505,485 tests, and the horizontal gold line is the suggestive threshold. (P-values for XY, Y, and mitochondrial SNPs are not shown.)

Table 15. Top 10 most associated SNPs from allelic association study. P-values are not corrected for multiple testing. Gene is in italics if SNP is not within the gene but near the gene. Genomic positions are on hg18.

Chr	Mb	SNP	Gene	A1	Freq AP	Freq Non-AP	ChiSq	P	OR
7	16.89	rs3735251	AGR3	G	0.314	0.099	28.81	7.98E-08	4.16
9	4.54	rs2039290	SLC1A1	A	0.086	0.008	28.35	1.01E-07	12.2
17	63.21	rs8065590	*NOL11/BPTF*	G	0.086	0.009	24.10	9.16E-07	10.0
13	39.78	rs2324591	*FOXO1*	C	0.243	0.076	21.77	3.07E-06	3.88
5	142.21	rs40127	ARHGAP26	G	0.300	0.108	21.70	3.20E-06	3.52
17	6.08	rs7503953	*WSCD1*	A	0.414	0.185	20.90	4.83E-06	3.12
15	52.56	rs1814785	UNC13C	G	0.400	0.117	20.61	5.63E-06	5.05
11	119.87	rs12364480	*ARHGEF12*	G	0.429	0.197	20.31	6.59E-06	3.05
8	0.80	rs2336409	BC022082	C	0.229	0.550	18.62	1.60E-05	0.24
4	96.95	rs6835311	*PDHA2*	A	0.457	0.227	18.47	1.72E-05	2.88

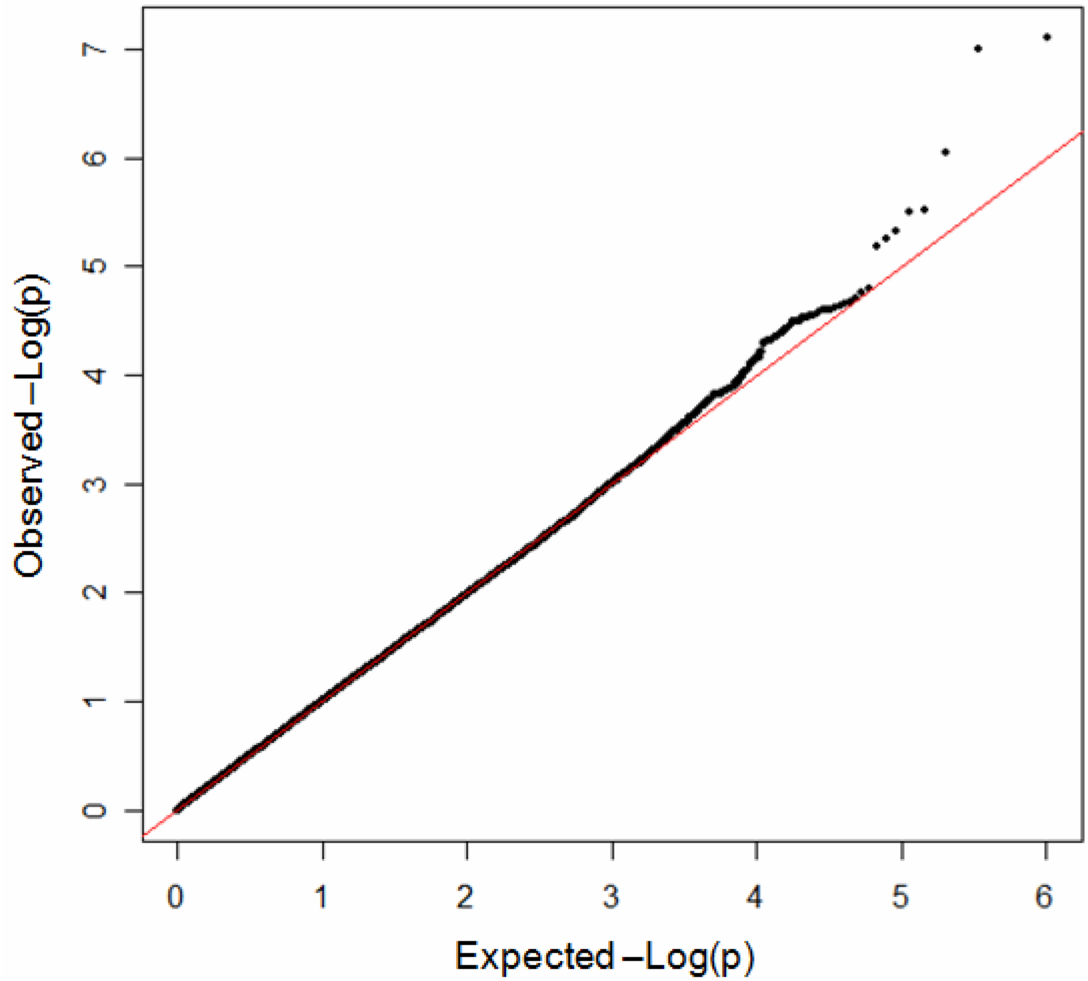


Figure 19. Q-Q Plot of final AJ allelic association results. Y axis is observed, X axis is expected –log(p).

Table 16. Follow-up case-control association study of five most significantly associated SNPs in 31 AP possessors and 24 non-AP possessors of non-Ashkenazi, European ancestry. Genomic positions are on hg19.

Chr	Bp	SNP	A1	Freq AP	Freq Non-AP	A2	# AP	# Non-AP	ChiSq	P	OR
5	142230251	rs40127	G	0.08065	0.02273	A	31	22	1.617	0.2036	3.772
7	16921334	rs3735251	G	0.04839	0.02273	A	31	22	0.4667	0.4945	2.186
9	4552653	rs2039290	T	0.01613	0.0625	C	31	24	1.66	0.1976	0.2459
13	40881549	rs2324591	G	0.2097	0.2292	T	31	24	0.06024	0.8061	0.8924
17	65776113	rs8065590	G	0.01613	0.0625	A	31	24	1.66	0.1976	0.2459

VII. CANDIDATE GENE RE-SEQUENCING AND ASSOCIATION

From the genome-wide linkage studies described in Chapters III-V, a handful of candidate linkage regions that might harbor genetic variants underlying AP emerged. We chose a subset of candidate genes within those regions to investigate by Sanger sequencing the exons and/or surrounding regions in about eight unrelated AP possessors of European ancestry. These decisions were informed by existing knowledge of where these genes are expressed, what their suspected functions are, and how close to the linkage peaks these genes were located. Due to time and financial constraints as well as overlap with the subsequent next-generation sequencing project (Chapter VIII), all of the exons were not sequenced in all of the genes we selected to pursue.

Our sequencing strategy included PCR amplification of genomic DNA or RT-PCR of RNA isolated from immortalized lymphoblastoid cell lines to analyze transcript sequences directly, and the resulting PCR products were Sanger sequenced at the UCSF Genomics Core Facility. Primer sequences are listed in Appendix B. Three summer interns, Androuw Carrasco, Yuri Cheung, and Ian McCulloch, assisted with the generation and analysis of sequencing data for many of the genes (Table 17). Overall, the majority of the single nucleotide variants we detected were already in dbSNP, and very few changed amino acids (Table 17, Appendix C).

For those SNPs that had been genotyped in CEU individuals for the HapMap project or had control allele frequencies for individuals of European ancestry from a different source, we compared the minor allele frequencies of the SNPs in AP individuals to those in CEU or other control individuals (Appendix C). Since SNPs in two genes, *ADCY8* and *TUSC1*, showed allele frequency differences between a subset of AP probands and CEU controls, we sequenced those genes more extensively (Tables 18-23).

ADCY8

As mentioned in Chapter IV, *adenylate cyclase 8* (*ADCY8*) is expressed almost exclusively in the brain[87] and is thought to play a role in learning and memory.[88,89] Specifically, the G allele of rs263249 in ADCY8 was found to be associated with greater episodic memory performance in humans.[89] When we sequenced the region surrounding this SNP in addition to other coding and non-coding portions of *ADCY8* in around 8 probands with AP, we found that certain alleles, including the G allele of rs263249, appeared to be enriched in our AP probands as compared to known HapMap CEU allele frequencies (Table 18).

We then sequenced additional AP individuals to gain genotype information about the SNPs that showed some initial signs of association as well as additional SNPs nearby. A block of SNPs near the 3' end of *ADCY8* continued showing signs of association following the increase in the number of AP individuals assayed (Table 18). This block of SNPs, and specifically the GT haplotype of two SNPs that showed early association (rs263249-rs873667), was in linkage disequilibrium with a 61 bp deletion just 3' of the gene (Table 19). We used the deletion to assay for the proposed AP-predisposing haplotype in additional cases and controls, but this examination no longer supported association. In fact, the deletion appeared to be at different frequencies in AP family probands and AP singletons with no reported family members with AP. The musically trained controls of European ancestry that we collected also appeared to have different deletion frequencies than the CEU individuals. When we verified the case-control deletion data by sequencing the SNP-containing region in LD with the deletion in some additional individuals, we got similar results (Table 20). We also looked at the segregation of the deletion in seven families of European descent (Figure 20). The deletion did not appear to clearly be associated with AP, as it was present in several individuals without AP and lacking in at least one individual with AP.

TUSC1

Tumor suppressor candidate 1 (*TUSC1*) was not a well studied gene when we chose to re-sequence it. Since it only consisted of one exon, it was known to be expressed in the brain,[102] and it was the closest gene to our chromosome 9p21 linkage peak, we decided it would be a good candidate for sequencing. A handful of SNPs in the gene had alleles that were enriched in AP individuals as compared to CEU individuals (Table 21).

Since the gene was only about 3 kb in length and there was a large degree of linkage disequilibrium encompassing it, we also investigated the haplotypes of all of the SNPs we found by sequencing the region. Upon inspection of our list of genotypes, there appeared to be three major haplotype groups present (Table 22). In the initial subset of individuals we sequenced, it appeared that group 2 haplotypes were associated with AP. However, as with the ADCY8 associations, the strength of the association diminished when additional cases and musically-trained controls were included (Table 23).

Table 17. Genes re-sequenced in whole or in part in AP individuals. Detailed information about the SNPs found is in Appendix C. The genes ADCY8 and TUSC1 were extensively sequenced, and the results are described in more detail subsequently. If a summer intern assisted with sequencing of a gene, their initials are listed. Coordinates are on hg19.

Chr	Mb	Gene	SNPs found	Intern
7	106.5	*PIK3CG*	1 synonymous and 1 in 3'UTR	AC
7	107.9	*NRCAM*	3 synonymous and 1 non-synonymous	AC
8	73.7	*KCNB2*	9 intronic SNPs	
8	73.9	*TERF1*	No exonic SNPs found	YC
8	74.6	*STAU2*	1 non-synonymous	YC
8	74.9	*TCEB1*	No exonic SNPs found	YC
8	75.3	*GDAP1*	5 in 3' UTR	YC
8	79.5	*PKIA*	No exonic SNPs found	YC
8	130.8	*GSDMC*	No exonic SNPs found	AC
8	130.9	*FAM49B*	1 just 3' of gene, 1 just 5' of gene	AC
8	131.2	*ASAP1*	1 nonsyn, 3 in 3'UTR, 3 just 5' of gene, 5 intronic	AC
8	131.9	*ADCY8*	2 in 5'UTR, 3 syn, 1 nonsyn, 63 intronic, 1 5', 11 3'	IM
8	133.0	*EFR3A*	1 nonsyn, 6 intronic, 4 in 3'UTR	IM
9	23.7	*ELAVL2*	3 in 3'UTR, 1 3' of gene	
9	25.7	*TUSC1*	10 in 3'UTR, 3 in 5'UTR, 4 5' of gene, 2 syn, 3 nonsyn	IM

Table 18. SNPs detected in *ADCY8* sequences that were tested for differences in allele frequencies between AP probands of European descent and HapMap CEU founders. Highlighted in yellow are SNPs that were sequenced in a larger number of AP probands due to promising association signals in the SNPs highlighted in blue in the initial set of probands. P-values less than 0.05 are shown in red for emphasis. Positions are on hg18 chromosome 8.

	SNP info		CEU		AP		Chisq		OR A1		
Bp	rs#	A1/A2	A1	A2	A1	A2	chisq	chidist p	OR	95%CI L	95% CI U
131848842	rs6470848	A/C	20	100	12	16	9.1892	**0.0024**	3.75	1.54	9.12
131849139	rs2572862	A/C	13	107	3	27	0.0175	0.8948	0.91	0.24	3.44
131849331	rs6990380	A/G	57	63	25	5	12.4350	**0.0004**	5.53	1.98	15.40
131849409	rs6990427	C/G	56	64	25	5	12.9898	**0.0003**	5.71	2.05	15.93
131849424	rs11997892	A/G	47	59	25	5	14.2706	**0.0002**	6.28	2.23	17.65
131849440	rs10097218	C/T	61	57	27	3	14.5594	**0.0001**	8.41	2.42	29.24
131852386	rs17225390	A/G	100	12	14	2	0.0458	0.8305	0.84	0.17	4.15
131861401	rs263247	C/T	52	68	2	12	4.3969	**0.0360**	0.22	0.05	1.02
131864442	rs263249	A/G	52	68	2	14	5.6063	**0.0179**	0.19	0.04	0.86
131864442	rs263249	A/G	52	68	8	38	9.6961	**0.0018**	0.28	0.12	0.64
131864593	rs873667	C/T	60	60	3	13	5.5447	**0.0185**	0.23	0.06	0.85
131864593	rs873667	C/T	60	60	9	37	12.6814	**0.0004**	0.24	0.11	0.55
131864671	rs873666	C/T	60	60	2	12	6.4324	**0.0112**	0.17	0.04	0.78
131864671	rs873666	C/T	60	60	8	36	13.4291	**0.0002**	0.22	0.10	0.52
131874785	rs11776881	A/C	86	34	15	17	6.9646	**0.0083**	0.35	0.16	0.78
131879758	rs13258256	C/T	98	22	20	10	3.2177	0.0728	0.45	0.18	1.09
131880021	rs7015079	G/T	72	48	28	2	12.0000	**0.0005**	9.33	2.12	41.01
131880835	rs16904360	G/T	11	109	5	25	1.4167	0.2339	1.98	0.63	6.22
131880862	rs16904361	A/T	109	11	25	5	1.4167	0.2339	0.50	0.16	1.58
131880957	rs12547373	A/G	109	11	25	5	1.4167	0.2339	0.50	0.16	1.58
131880998	rs12545113	G/T	19	101	5	25	0.0124	0.9113	1.06	0.36	3.12
131886873	rs17226545	C/T	69	39	12	4	0.7595	0.3835	1.70	0.51	5.62
131886953	rs384271	A/G	81	39	13	3	1.2505	0.2635	2.09	0.56	7.75
131886957	rs402620	C/T	32	88	4	12	0.0201	0.8871	0.92	0.28	3.05
131887071	rs1543020	G/T	14	106	2	14	0.0094	0.9226	1.08	0.22	5.27
131895803	rs1435446	A/G	25	95	3	13	0.0375	0.8465	0.88	0.23	3.32
131895882	rs263265	A/G	25	95	1	15	1.9417	0.1635	0.25	0.03	2.01
131896244	rs4736704	C/T	95	25	13	3	0.0375	0.8465	1.14	0.30	4.31
131896313	rs377711	C/G	7	105	1	15	0.0000	1.0000	1.00	0.11	8.71
131905302	rs6996688	C/T	92	28	12	4	0.0218	0.8826	0.91	0.27	3.06
131905535	rs6997439	G/T	92	28	12	4	0.0218	0.8826	0.91	0.27	3.06
131905801	rs263263	A/T	97	23	13	3	0.0016	0.9682	1.03	0.27	3.91
131905912	rs263264	A/G	85	35	14	2	1.9802	0.1594	2.88	0.62	13.35
131906087	rs12375420	C/T	99	21	10	6	3.5490	0.0596	0.35	0.12	1.08
131922543	rs6993838	C/G	7	113	1	15	0.0044	0.9470	1.08	0.12	9.36
131922640	rs263256	A/T	97	23	13	3	0.0016	0.9682	1.03	0.27	3.91
131924906	rs263258	A/G	85	35	14	2	1.9802	0.1594	2.88	0.62	13.35
131925225	rs17227830	A/G	11	109	5	11	6.6324	**0.0100**	4.50	1.32	15.34
131925225	rs17227830	A/G	11	109	8	38	2.2193	0.1363	2.1	0.78	5.57
131925252	rs263260	A/G	85	35	14	2	1.9802	0.1594	2.88	0.62	13.35
131925303	rs16904374	A/G	25	95	4	12	0.1461	0.7023	1.27	0.38	4.27
131929803	rs12543363	A/G	92	28	12	4	0.0218	0.8826	0.91	0.27	3.06
131929989	rs12548296	C/T	83	35	11	5	0.0170	0.8963	0.93	0.30	2.87
131929994	rs7820412	G/T	33	69	5	11	0.0077	0.9300	0.95	0.31	2.96
131930184	rs12548835	C/T	85	35	11	5	0.0295	0.8636	0.91	0.29	2.80
131948841	rs12544368	C/T	72	48	6	10	2.9220	0.0874	0.40	0.14	1.17
131966264	rs4128982	C/T	88	32	14	2	1.5111	0.2190	2.55	0.55	11.82
131991209	rs12545028	G/T	15	105	3	13	0.4802	0.4883	1.62	0.41	6.34
132071758	rs1329803	A/G	70	50	7	9	1.2224	0.2689	0.56	0.19	1.59
132122594	rs913818	A/G	8	112	1	15	0.0040	0.9498	0.93	0.11	7.99
132123334	rs3829210	C/T	55	65	9	7	0.6149	0.4330	1.52	0.53	4.35

Table 19. Genotypes of AP family probands, AP singletons, musically trained individuals without AP, and CEU individuals for 61 bp deletion just 5' of *ADCY8* (rs55861470). The deletion was completely correlated with the GT haplotype (rs263249-rs873667) in all of the samples we investigated. The deletion was not significantly associated with AP whether CEU individuals were included as controls (χ^2=3.067, p=0.08) or not (χ^2=0.0007, p=0.97).

	+/+	+/-	-/-	Freq without deletion
AP probands	1	7	15	19.6%
AP singletons	5	19	7	46.8%
Non-AP controls	4	9	11	35.4%
CEU	14	32	14	50.0%
All AP	6	26	22	35.2%
All without known AP	18	41	25	45.8%

Table 20. Comparison of allele frequencies of *ADCY8* SNPs in AP probands of European descent to those in musically trained controls without AP of European descent. Positions are on hg18.

SNP info			Controls		AP		Chisq		OR A1		
Bp	rs#	A1/A2	A1	A2	A1	A2	chisq	chidist p	OR(A1)	95%CI L	95% CI U
131864442	rs263249	A/G	5	21	8	38	0.038	0.845	0.884	0.256	3.049
131864593	rs873667	C/T	7	19	9	37	0.520	0.471	0.660	0.213	2.048

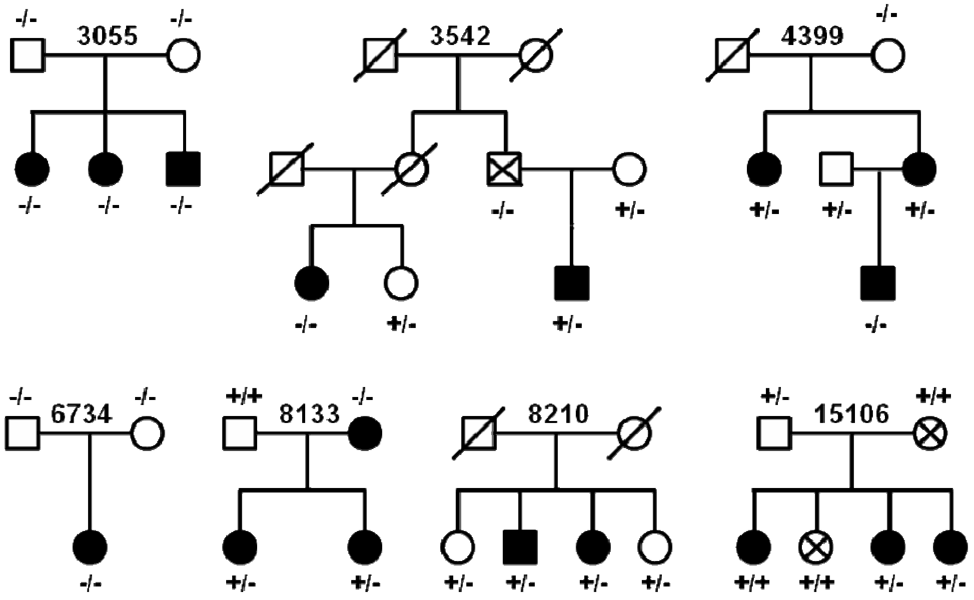


Figure 20. Segregation of *ADCY8* 3' deletion in seven families of European descent.

Table 21. *TUSC1* SNP allele frequency differences between AP cases and CEU controls of unknown AP status. CEU data highlighted in yellow was obtained from sequencing a subset of HapMap CEU samples. P-values less than 0.05 are shown in red. Physical positions are on hg18 chromosome 9.

SNP info			CEU		AP		Chisq		OR A1		
Bp	rs#	A1/A2	A1	A2	A1	A2	chisq	chidist p	OR	95%CI L	95% CI U
25666955	rs7028310	C/G	101	19	37	5	0.380	0.537	1.392	0.485	3.997
25667215	rs12348	C/T	56	64	12	28	3.410	0.065	0.490	0.228	1.053
25667257	rs1128957	C/G	70	50	16	24	4.056	0.044	0.476	0.230	0.987
25667349	rs1128953	G/T	71	49	20	24	2.451	0.117	0.575	0.287	1.154
25667588	rs10812300	G/T	22	34	31	19	5.451	0.020	2.522	1.152	5.519
25667698	rs72631815	G/T	13	43	24	26	7.142	0.008	3.053	1.328	7.018
25667933	rs35110225	A/G	28	28	18	30	1.637	0.201	0.600	0.274	1.315
25667953	rs34498078	A/G	2	54	4	46	0.970	0.325	2.348	0.411	13.408
25668122	rs72631814	C/G	47	11	26	18	5.921	0.015	0.338	0.139	0.823
25669024	rs10738727	C/G	73	47	16	26	6.497	0.011	0.396	0.192	0.816

Table 22. Major *TUSC1* haplotypes observed in sequenced individuals. Physical positions are on hg18 chromosome 9.

Bp	SNP rs#	1	1.1	1.2	2	2.1	2.2	2.3	2.4	3
25666579	N/A	G	G	G	G	G	G	G	G	G
25666584	rs4592123	G	G	G	G	G	G	G	G	T
25666809	rs10812298	G	G	G	T	T	T	T	T	T
25666895	rs10812299	G	G	G	A	A	A	A	A	A
25666913	rs7044566	T	T	T	A	A	A	A	A	T
25666955	rs7028310	G	G	G	G	G	G	G	G	C
25667215	rs12348	G	G	G	A	A	A	A	A	A
25667257	rs1128957	C	C	C	G	G	G	G	G	C
25667349	rs1128953	G	G	G	T	T	T	T	T	G
25667588	rs10812300	T	T	T	G	G	G	G	T	G
25667698	rs72631815	T	T	T	G	G	G	T	T	T
25667933	rs35110225	A	A	G	G	G	G	G	A	G
25667953	rs34498078	G	G	G	G	G	G	G	A	G
25668122	rs72631814	C	C	C	G	G	C	C	C	C
25668196	rs72631813	A	G	A	A	A	A	?	A	A
25668639	rs61483294	C	C	C	C	C	C	T	C	C
25668640	rs10967034	T	T	T	A	A	A	A	A	A
25668797	rs34772164	C	C	C	C	C	C	C	C	T
25668887	rs60018547	C	C	C	T	C	T	T	C	C
25669024	rs10738727	G	G	G	C	C	C	C	C	G
25669137	rs10738728	A	A	A	G	G	G	G	G	A
25669141	rs10738729	A	A	A	G	G	G	G	G	A

Table 23. *TUSC1* haplotype distributions in individuals with and without AP. Haplotype group 2 was not significantly associated with AP (χ^2=1.61, p=0.20) when only haplotype groups 1 and 2 were included.

	1	2	3	Other	2/(1+2)
AP probands	13	29	5	3	69.0%
AP singletons	12	5	3	0	29.4%
Non-AP controls	5	15	0	0	75.0%
CEU	32	17	8	1	34.7%
All AP	25	34	8	3	57.6%
All without known AP	37	32	8	1	46.4%

VIII. TARGETED NEXT-GENERATION SEQUENCING OF GENES IN LINKAGE REGIONS

Following the SNP linkage study described in Chapter IV, four broad linkage regions were identified in which to search for genetic variants that may influence the development of AP. As described in Chapter VII, our attempts to find these variants by Sanger sequencing good candidate genes (based on their expression pattern, proposed function, or genomic location) resulted in few promising leads. Several options for follow-up were then considered, such as fine mapping the region with additional markers to narrow the linkage regions of interest and to look for the association of particular marker alleles with AP, adding additional individuals to increase the power of our study, or continuing to Sanger sequence candidate genes. However, recent advances in targeted next-generation sequencing technology prompted us to pursue that new approach, since it promised to give us a wealth of genetic data for a reasonable price. Specifically, we opted to sequence the majority of the genes in our four linkage regions in ten unrelated AP possessors by first capturing our target sequences and then sequencing the population of selected molecules.

Target selection

To capture DNA fragments from our regions of interest, we chose the Agilent SureSelect Target Enrichment System[103] due to its reportedly high specificity and its ability to pair well with next-generation sequencing on the Illumina Genome Analyzer, the most cost-effective sequencing option for our project at the time. Using Agilent's eArray software, we designed a total of 57,678 unique 120-mer baits covering approximately 3.84 Mb of the human genome using a 2x tiling approach. After masking repeats, we were able to cover nearly all of the intronic and exonic portions of genes in the chromosome 8q24, 8q21, 7, and 9 linkage regions plus 10 kilobases upstream and downstream of each gene (Figure 21). In fact, we were able to include all non-repeat base pairs in the 8q24 linkage region between *GSDMC* and *ADCY8*. We also included a few non-protein-coding regions of interest, including an enhancer and several potential small RNA encoding genes.

Library preparation

Since we aimed for an average of 30-40X sequence read depth on the 3.84 Mb of sequence that we targeted and since we chose to sequence the targeted regions in 10 unrelated AP possessors, we prepared multiplexed libraries so that sequencing reads from more than one sample in a single lane of the Genome Analyzer could be distinguished from one another. We also chose to create paired-end libraries to enable reads to be mapped to the genome with greater confidence and to gain some copy number variation information from our sequencing data.

To prepare the genomic DNA for library construction, 10 micrograms of each of the 10 genomic DNA samples were sonicated and run on separate 2% agarose gels using Biorad Certified Low Range Agarose in TAE alongside an NEB Low Molecular Weight Ladder. DNA fragments ranging from 200 to 300 bp were extracted from the gel. Using Illumina kit reagents from the paired-end sample preparation kit and multiplexing kit, the ends were repaired, an A overhang was added to one end, and the Index Paired End Adapters were ligated onto the ends of the fragments (Figure 22). The samples were then run on another 2% gel, and DNA fragments that were approximately 300 bp long were purified from a 1-2 mm slice of the gel. Six cycles of PCR were performed using the Index PE PCR primer 1.0, Index PE PCR primer 2.0, and unique index primers for each sample to amplify the libraries and add the sequences necessary for cluster generation and sequencing of the samples onto the end of the library fragments. These samples were purified using the Qiagen PCR Purification Kit.

We then used the reagents provided in the Agilent SureSelect Target Enrichment System to hybridize the custom baits to our library fragments overnight, using the PE block reagent. (It should be noted that Agilent had not yet developed a blocking reagent for the Multiplex PE primers, so we used the non-multiplex PE ones as a compromise, which could have resulted in some reduction in specificity.) We retrieved DNA hybridized to the biotin-labeled baits using magnetic streptavidin beads. An additional 12 cycles of PCR using 7 µl of the captured library, Herculase II Fusion DNA Polymerase, the Index PE PCR primer 1.0, and the unique index primers were performed, and the resulting PCR products were purified using a Qiagen PCR Purification Kit.

Sanger sequence of clones

Prior to next-generation sequencing, the success and specificity of the library preparation and capture was assayed by cloning and Sanger sequencing some of the library fragments. Two microliters of each resulting library were cloned into a TOPO blunt vector, plated, and 10 colonies were grown up for Sanger sequencing. In total, 99 library fragments were cloned and Sanger sequenced for library validation. (One of 100 did not sequence properly.) When the DNA sequences from these fragments were mapped to the human genome using BLAST, 87 of 99 (87.9%) mapped in or near the targeted candidate regions on chromosomes 7, 8, and 9. The majority of the sequences matched perfectly with the reference genome, but there were also 24 single nucleotide variants detected, 21 of which were known SNPs in dbSNP.

In contrast, the oligonucleotide sequences that had been added onto the ends of the fragments as adapters or incorporated into PCR primers did not exhibit very high fidelity in the sequenced clones. Only 54 of 99 clones had the expected adapter/primer length and sequence (Figure 22) on both ends. 14 of 99 had the correct length but had mismatches within the adapter/primer sequences, with 3 of these having 2 different mismatches. 30 of the 99 had truncated adapter/primer sequences on one end, and 5 of these also had mismatches within the adapter/primer sequences. Finally, one cloned fragment had truncations on both ends of the DNA sequence. A lot of the truncations were not small (Table 24), so it was expected that they could prevent the fragments from binding to complementary sequences on the flow cell during the cluster generation step of sequencing. Thus, the concentration of sequenceable library fragments was lower than the DNA concentration of the library itself.

We were also able to determine the average library insert size from the Sanger sequences of our clones. Since we cut out bands around 300 bp after adapters were ligated but before PCR with primers that added additional bp onto our fragments was performed, we expected the average length of our cloned fragments to be an average of 368 bp, including adapter/primer sequences. The distribution of clone lengths from our 10 libraries is shown in Figure 23, and the average length was 365.1 bp.

Library quantitation and sequencing

Our libraries were sequenced on three different occasions. On the first attempt in January 2010, libraries were quantified using a NanoDrop spectrophotometer, pooled, and diluted according to measured NanoDrop concentrations. Three to four libraries were pooled into one tube so that all 10 libraries could be sequenced on three lanes of an Illumina Genome Analyzer II flow cell. Based on the DNA concentrations measured on the NanoDrop, 5 pM of DNA was loaded onto each of three lanes of the flow cell. The samples were then multiplex paired-end sequenced with 65 bp in each direction in addition to the 6 bp multiplex index read at the UCSF Center for Advanced Technology (CAT) with the help of Clement Chu. Unfortunately, the sequence yield proved to be about 40-fold less than expected, with only 3,000-5,000 clusters/tile instead of the expected 130,000-140,000 clusters/tile.

Before the second sequencing attempt in April 2010, we first compared the sequenceable molecule concentrations of our samples to a control using quantitative real-time PCR with SYBR green, in an attempt to get a more accurate estimate of the library concentration. We used this information from each of our libraries to pool all 10 into one tube at approximately equal concentrations. In addition to our own qPCR analyses, library DNA concentrations were quantified using picogreen and in-house qPCR by Leath Tonkin at the QB3/UC Berkeley Genomics Sequencing Laboratory (GSL). He reported a 9.88 nM concentration by picogreen and 2.85 nM by qPCR. Based on the qPCR estimate, 6 pM of the 10 pooled libraries was added to one lane of the flow cell, and 8 pM was added to a second lane of the flow cell. The multiplexed paired end reads were 76 bp in each direction this time. The yield was a great improvement over the previous attempt, with approximately 60,000 clusters/tile in the first lane and 70,000 clusters/tile in the second lane.

Since these yields were still about 4-fold less than expected, we had the library sequenced one more time in one lane at the GSL. On the third and final attempt in May 2010, 5-fold more of the library was added to one lane than the April 2010 attempt, but there were still only about 90,000 clusters/tile (the maximum possible was about 300,000 clusters/tile). Perhaps the library was nuclease contaminated, which would explain why the effective concentration of the library decreased on successive sequencing attempts. Regardless of the reason, sufficient sequence data had been generated for adequate coverage of the majority of our targeted regions, so we proceeded with data analysis.

SNP and indel discovery

The first step of data analysis was to use the Illumina software CASAVA v1.6 to de-multiplex the pooled sequences based on their 6-bp index sequences into separate files. The resulting sequence files were then mapped to hg19 using the Burrows-Wheeler Alignment tool (BWA).[104] Once mapped, the files in sequence alignment/map (SAM) format were imported into SAMtools[105] for SNP and indel calling. Sequence reads from January, April, and May were combined into one file for each of the 10 sequenced samples. Reads that had identical sequences on both ends were removed, because they were likely PCR duplicates. In all, we had data from over 54 million reads, which totaled just over 4 Gb of sequence (Table 25). This gave us adequate coverage in the majority of our targeted regions (Figure 24).

Before applying any stringent filtering criteria, we detected a number of small differences from the hg19 reference sequence in our 10 samples (Figure 25). These were classified as single nucleotide polymorphisms (SNPs) or insertion/deletion variants (indels) and further subdivided based on their location, novelty, and other characteristics.

Many SNP non-reference alleles were found in only one sample, but a number were found in multiple samples, either as heterozygotes or homozygotes (Figure 26). As expected, SNPs that were not already in dbSNP were less common than those that were in dbSNP, and SNPs in exons were less common than SNPs outside of exons on average. All of the exonic indels were in untranslated regions, leaving their functional significance open to question.

Association of known SNPs with AP

Since we hypothesized that AP arises from a combination of genetic and environmental influences, it was possible that previously described variants (i.e. those within dbSNP) could be AP-predisposing variants, so we decided to test this using an association study. Though we did not sequence any control individuals in our study, we could still compare allele frequencies of the SNPs we detected in our 10 AP individuals to control allele frequencies in publicly available databases.

Of the SNPs we detected, 3,070 were genotyped in the HapMap phase III project,[106,107] enabling us to compare the AP allele frequencies to HapMap CEU allele frequencies for those variants (Figure 27, Table 26). In addition, the 1000 genomes phase I project[108] sequenced CEU individuals at low coverage over the majority of the genome, allowing us to garner some approximate control allele frequencies for 10,822 SNPs from those data for a second allelic association study (Figure 28, Table 27). All p-values shown were uncorrected for multiple testing; p-values that more accurately reflect the significance of the associations could be obtained by multiplying by the number of genes tested. The number of genes we looked at fell between 70 and 110, depending on whether hypothetical genes, pseudogenes, and non-protein-coding genes are included.

We looked at allele frequencies reported in dbSNP for the SNPs that appeared to be most associated with AP, and sometimes if the frequencies reported in dbSNP were substituted for the CEU data from HapMap or 1,000 genomes, the strength of the association decreased, so we did not consider those SNPs to be strong candidates. Fewer than 1,000 SNPs already in dbSNP were not included in at least one of these allelic association studies, so they would have to be genotyped in control individuals to determine if they are associated with AP.

Novel SNPs

Since AP is rare, it was possible that some or all AP-predisposing variants could also be rare enough to not have been entered into dbSNP yet. The majority of SNPs that we detected in our next-generation sequencing data that were not in dbSNP were seen in only one individual. Of the 13 non-synonymous SNPs detected, 3 had low SNP quality scores so they may not have been real, and none of the 10 remaining non-synonymous SNPs were in the same gene (Table 28). Thus, novel non-synonymous variants may not explain the majority of AP genetic predisposition. Non-coding variation could also play a role in AP, so we took note of novel variants that were seen in multiple AP individuals even though their functional consequences may be more difficult to determine (Table 29).

Copy-number variation

Since we obtained sequence data from both ends of library DNA fragments, we were able to determine if there were any outliers in the distance between the ends when mapped to the genome, thus indicating the potential presence of a long indel or a CNV. Some of the outliers with ends that mapped closer to each other than expected appeared to result from mis-priming by the sequencing primer, because the genomic sequence just upstream of the read was complementary to the 3' end of the primer used (data not shown). We detected six large previously described deletions in our 10 AP individuals (Table 30). We assayed four of these in additional AP and non-AP possessors of European descent, but none showed promising signs of association with AP after these analyses. We did not assay the LAMB4 deletion because it was difficult to design unique primers in the area due to repeats, and we did not assay the EYA1 deletion because there was already a known control frequency that seemed similar to what we observed in AP individuals.

Follow-up association study in additional Eu individuals

To follow-up on some of the SNPs we detected in our sequence from 10 AP possessors in additional AP possessors and non-AP possessors of European descent, we employed the Sequenom iPlex Gold Genotyping platform because it was the most cost-effective option for the number of samples and SNPs we wanted to genotype. Sequenom assay design and genotyping was carried out by the UCSF Cancer Center Genome Analysis Core.

Our wish list for follow-up genotyping included novel non-synonymous SNPs, the majority of novel non-exonic SNPs seen in at least 3 different individuals that were not in obvious linkage disequilibrium with one another, and selected SNPs that were already in dbSNP that showed promising evidence for association when HapMap and/or 1,000 genomes project CEU genotype data were used as controls. Unfortunately, Sequenom assays could not be successfully designed for some of the SNPs we were interested to pursue, but we were able to order assays for 52 different SNPs. Six of the SNPs for which assays were designed failed to genotype on the Sequenom platform and another 4 appeared monomorphic because only one of two alleles was ever observed, making those SNPs useless for association analyses. Three rare SNPs were only seen in the individuals that were originally sequenced and their family members, and another four SNPs were seen in multiple AP families but no singleton AP cases or controls.

We first used 31 AP cases and 24 non-AP controls to determine if the 27 SNPs that were informative in those individuals were associated with AP (Table 31). Unfortunately, no significant association was detected using that relatively small population. We also were able to genotype some individuals from 48 of our families of European descent for use in a combined association study using DFAM in Plink,[82] which includes the sibling transmission disequilibrium test (sib-TDT) and also incorporates genotype data from unrelated individuals. One of the 28 SNPs that were informative had a p-value less that 0.05 in this analysis, but this association would not be significant after correction for multiple testing (Table 32.) Overall, a larger sample size may be necessary to adequately determine whether the SNPs that were assayed play a role in AP, but it is also possible that we need to look elsewhere for AP-predisposing variants.

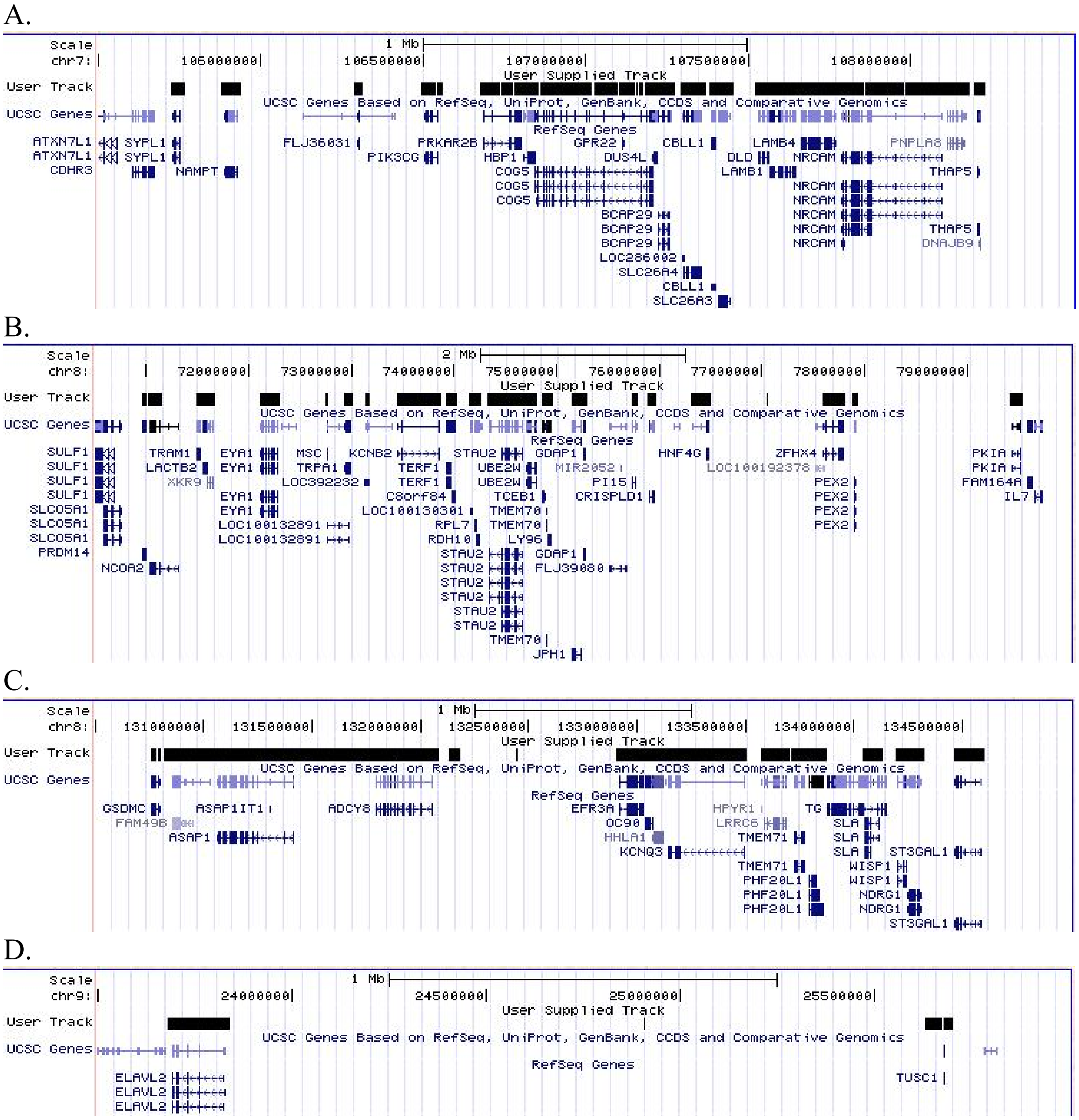


Figure 21. Genomic regions targeted for next-generation sequencing are shown in the user track as black bars on (A) chromosome 7 (B) chromosome 8q21 (C) chromosome 8q24 and (D) chromosome 9.

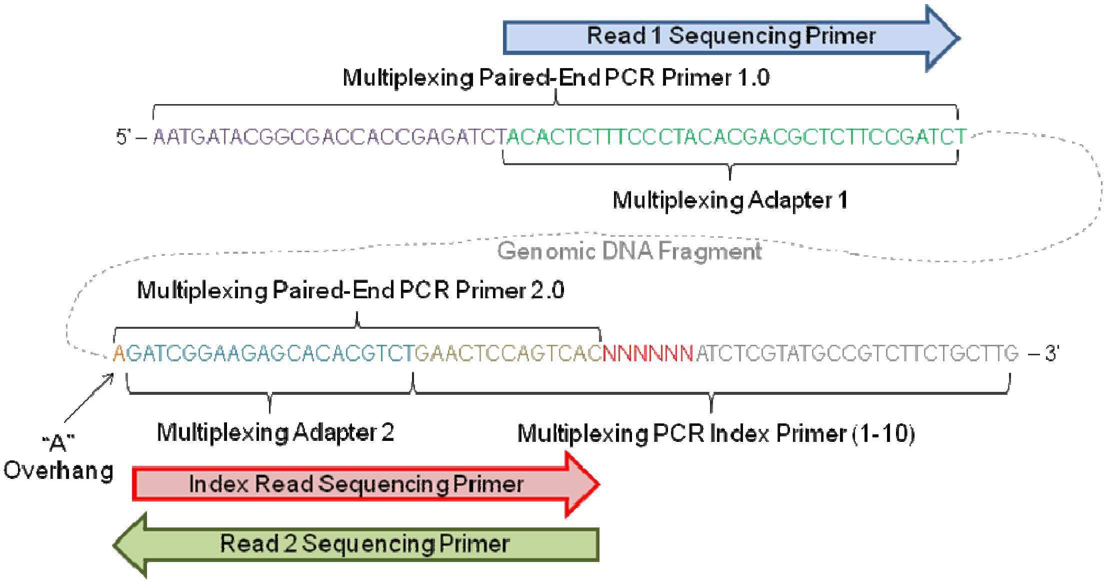

Read 1 Sequencing Primer
Multiplexing Paired-End PCR Primer 1.0
5' – AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Multiplexing Adapter 1
Genomic DNA Fragment
Multiplexing Paired-End PCR Primer 2.0
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG – 3'
"A" Overhang
Multiplexing Adapter 2
Multiplexing PCR Index Primer (1-10)
Index Read Sequencing Primer
Read 2 Sequencing Primer

Figure 22. Expected anatomy of a multiplexed paired-end library fragment.

Table 24. Truncations observed in 31 of 99 cloned library fragments. One fragment had both ends truncated.

Size of truncation	# of truncated ends
1 bp	3
2 bp	10
3 bp	6
4 bp	1
5 bp	1
6 bp	1
7 bp	3
8 bp	1
9 bp	5
10 bp	1
Total	32

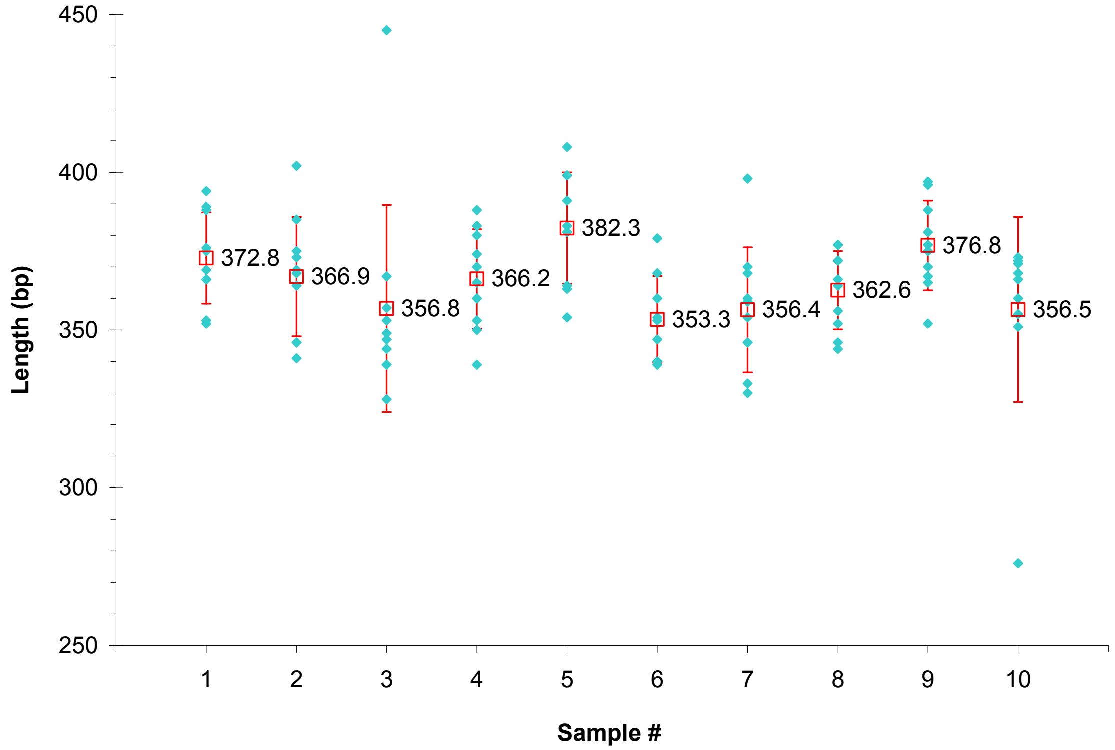


Figure 23. Distribution of cloned library fragment lengths including adapter/primer sequences. The average length across all libraries was 365.1 bp.

Table 25. Numbers of mapped reads for each sample from each sequencing run (January, April, and May) after removal of probable PCR duplicate reads. Based on these numbers, the total basepairs of mapped sequence were calculated.

Index	JanUnpair	JanPair	AprUnpair	AprPair	MayUnpair	MayPair	TotalUnpair	TotalPair	TotalBp
1	5984	69974	18672	3348272	6110	1936810	30766	5355056	405864672
2	4344	42604	15644	2915392	6626	2064414	26614	5022410	381021496
3	5692	52506	19244	3531412	9152	2628766	34088	6212684	471323888
4	5704	65222	17656	3173920	7950	2403335	31310	5642477	427784700
5	9372	104750	38170	4249178	18486	2828562	66028	7182490	544193240
6	15632	142892	20138	3479782	9950	2701337	45720	6324011	478338564
7	3946	37874	14010	2446572	5608	1765867	23564	4250313	322417804
8	8222	72666	17926	3009462	8510	2263696	34658	5345824	405119968
9	2390	28348	15650	2569598	5772	1431722	23812	4029668	305801200
10	8924	81560	17660	2733654	8622	2026076	35206	4841290	366633080
# All	70210	698396	194770	31457242	86786	22050585	351766	54206223	4108498612

# #

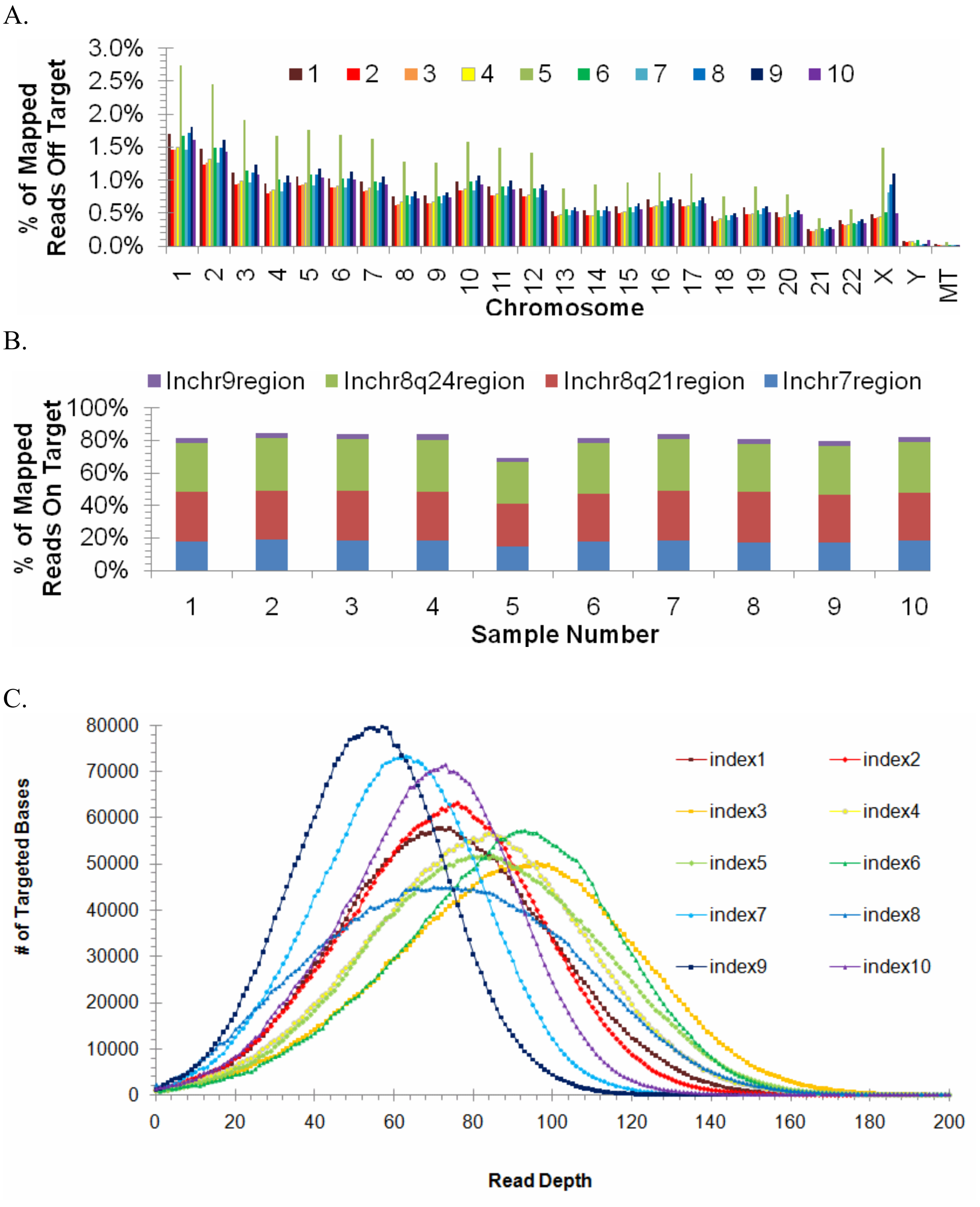


Figure 24. Targeting specificity and read depth in targeted regions. (A) Proportion of reads that mapped off target in each sample, (B) proportion of reads that mapped to targeted linkage regions in each sample, and (C) sequencing read depth of targeted regions in each sample.

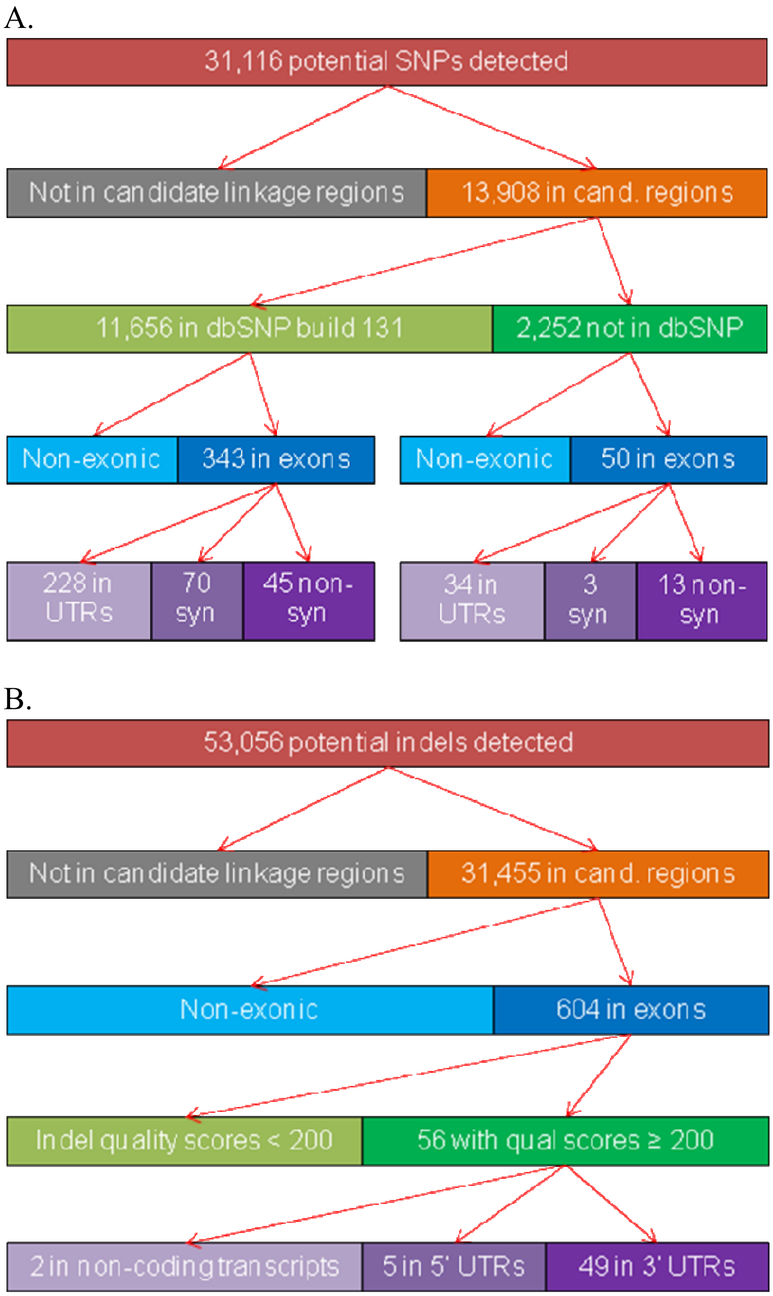

A.
31,116 potential SNPs detected
Not in candidate linkage regions
13,908 in cand. regions
11,656 in dbSNP build 131
2,252 not in dbSNP
Non-exonic
343 in exons
Non-exonic
50 in exons
228 in UTRs
70 syn
45 non-syn
34 in UTRs
3 syn
13 non-syn
B.
53,056 potential indels detected
Not in candidate linkage regions
31,455 in cand. regions
Non-exonic
604 in exons
Indel quality scores < 200
56 with qual scores ≥ 200
2 in non-coding transcripts
5 in 5' UTRs
49 in 3' UTRs

Figure 25. Potential (A) SNPs and (B) indels detected in next-generation sequencing data.

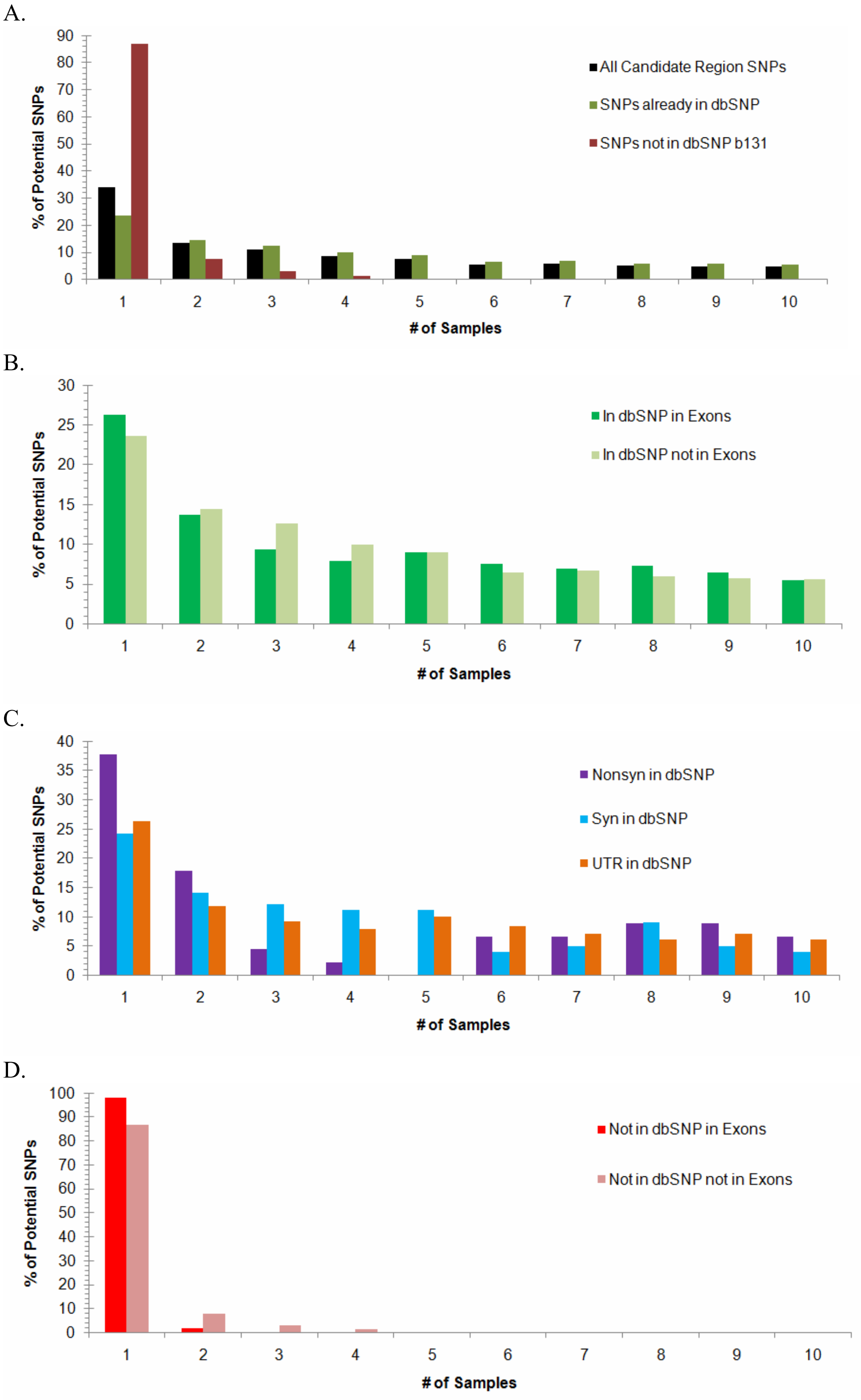


Figure 26. The percentage of potential SNPs versus the number of samples they were detected in for (A) SNPs in chromosome 7, 8, and 9 candidate regions (B) SNPs in dbSNP build 131 (C) Exonic SNPs in dbSNP and (D) SNPs not in dbSNP build 131.

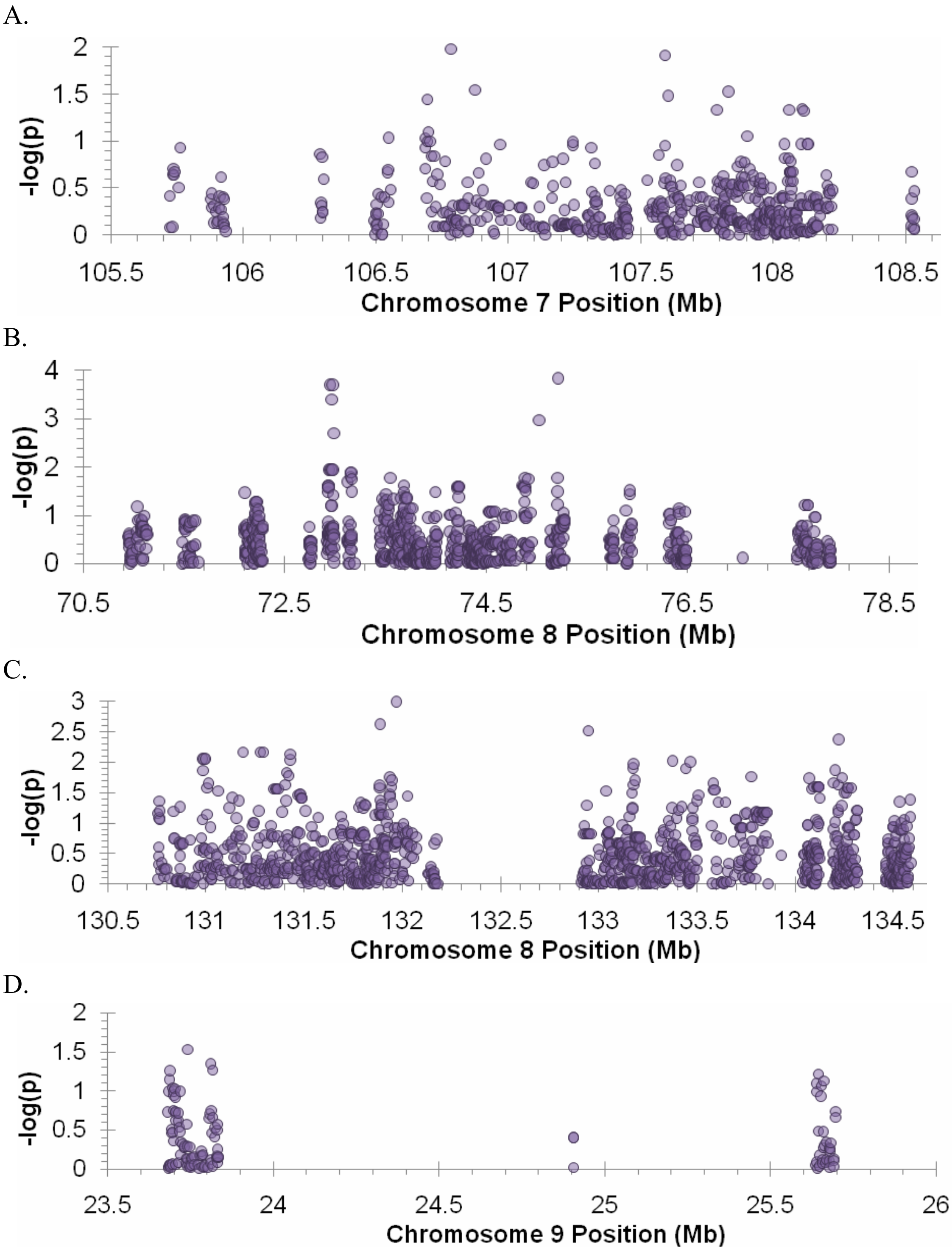


Figure 27. Allelic association study using 10 sequenced AP individuals as cases and CEU individuals genotyped for the HapMap project as controls. Results are shown for the (A) chromosome 7 (B) chromosome 8q21 (C) chromosome 8q24 and (D) chromosome 9 linkage regions. Physical positions are on hg19. P-values are not corrected for multiple testing.

Table 26. Top 50 most associated SNPs when using 10 sequenced AP possessors as cases and 113 CEU individuals genotyped for the HapMap project as controls. P-values are not corrected for multiple testing. Positions are on hg19.

SNP info						CEU		AP		Statistics	
Chr	Bp	rsName	Gene	Ref	Alt	#ref	#tot	#ref	#tot	ChiSq	p
7	106783164	rs12705404	*PRKAR2B*	T	C	57	226	0	20	6.566	0.01040
7	107592198	rs2072209	*LAMB1*	A	G	208	226	15	20	6.291	0.01213
8	72941733	rs3824150	*TRPA1*	A	T	163	226	9	20	6.427	0.01124
8	72953158	rs1025926	*TRPA1*	C	T	183	226	9	20	13.879	0.00020
8	72957716	rs1373302	*TRPA1*	T	A	163	226	9	20	6.427	0.01124
8	72965973	rs3735942	*TRPA1*	G	A	163	226	9	20	6.427	0.01124
8	72969263	rs3779752	*TRPA1*	A	C	200	226	12	20	12.526	0.00040
8	72976997	rs16937961	*TRPA1*	C	T	183	226	9	20	13.879	0.00020
8	72978593	rs7824377	*TRPA1*	C	T	163	226	9	20	6.427	0.01124
8	72980652	rs1443952	*TRPA1*	C	T	163	226	9	20	6.427	0.01124
8	72987384	rs2278653	*TRPA1*	A	T	181	226	10	20	9.582	0.00196
8	73122967	rs10112844	*LOC392232*	C	T	129	226	6	20	5.441	0.01967
8	73156028	rs4324960	*AK309726*	T	G	131	226	6	20	5.823	0.01582
8	73158374	rs1482133	*AK309726*	G	C	133	226	6	20	6.222	0.01261
8	73162530	rs4738225	*AK309726*	C	A	133	226	6	20	6.222	0.01261
8	73164823	rs13259433	*AK309726*	A	G	130	226	6	20	5.630	0.01766
8	73164998	rs13260014	*AK309726*	A	G	130	226	6	20	5.630	0.01766
8	73548109	rs1489221	*KCNB2*	C	T	207	226	15	20	5.746	0.01653
8	74884530	rs6990813	*TCEB1*	G	T	137	216	18	20	5.734	0.01663
8	74918871	rs1426060	*LY96*	T	C	144	226	18	20	5.645	0.01751
8	75207653	rs4738443	*JPH1*	G	A	91	214	3	20	5.765	0.01635
8	75213649	rs16938860	*JPH1*	C	T	223	226	17	20	14.435	0.00015
8	130982439	rs16904182	*FAM49B*	C	T	209	226	15	20	6.893	0.00866
8	130983869	rs16904183	*FAM49B*	A	G	204	222	15	20	6.087	0.01362
8	130990662	rs17279655	*FAM49B*	T	C	209	226	15	20	6.893	0.00866
8	130996692	rs17194770	*FAM49B*	C	T	209	226	15	20	6.893	0.00866
8	131189298	rs11777289	*ASAP1*	T	C	215	226	16	20	7.349	0.00671
8	131276243	rs7826256	*ASAP1*	C	T	215	226	16	20	7.349	0.00671
8	131291187	rs11778881	*ASAP1*	T	C	215	226	16	20	7.349	0.00671
8	131409885	rs4609234	*ASAP1*	A	G	142	226	7	20	5.959	0.01464
8	131421581	rs5027392	*ASAP1*	G	A	141	226	7	20	5.751	0.01648
8	131429723	rs3924865	*ASAP1*	T	A	124	226	17	20	6.819	0.00902
8	131430133	rs7386870	*ASAP1*	T	C	121	224	17	20	7.173	0.00740
8	131885942	rs12155610	*ADCY8*	C	T	117	226	16	18	9.264	0.00234
8	131936854	rs6470861	*ADCY8*	C	A	95	224	14	20	5.654	0.01741
8	131945428	rs6997554	*ADCY8*	T	C	132	226	17	20	5.441	0.01967
8	131970181	rs6470872	*ADCY8*	A	G	2	226	2	18	10.812	0.00101
8	132947786	rs16904553	*EFR3A*	C	T	180	218	11	20	8.786	0.00304
8	133174006	rs7007544	*KCNQ3*	T	G	193	224	13	20	6.253	0.01240
8	133176146	rs16904609	*KCNQ3*	G	A	199	226	12	18	6.520	0.01067
8	133185294	rs6991887	*KCNQ3*	C	T	198	224	14	20	5.451	0.01956
8	133377440	rs2100646	*KCNQ3*	C	T	181	226	11	20	6.750	0.00937
8	133444342	rs2673593	*KCNQ3*	T	G	171	226	10	20	6.225	0.01260
8	133466893	rs3857927	*KCNQ3*	G	T	219	226	17	20	6.675	0.00978
8	133776694	rs2553603	*TMEM71*	T	C	20	226	4	14	5.697	0.01699
8	134071833	rs2741200	*TG/SLA*	T	C	159	226	19	20	5.580	0.01817
8	134196037	rs4736640	*WISP1*	T	G	176	226	11	20	5.274	0.0216
8	134202942	rs2013158	*WISP1*	C	A	179	226	11	20	6.122	0.01335
8	134220691	rs2929946	*WISP1*	A	G	89	226	1	18	8.194	0.00420
8	134240697	rs2929969	*WISP1*	G	A	33	226	7	20	5.615	0.01781

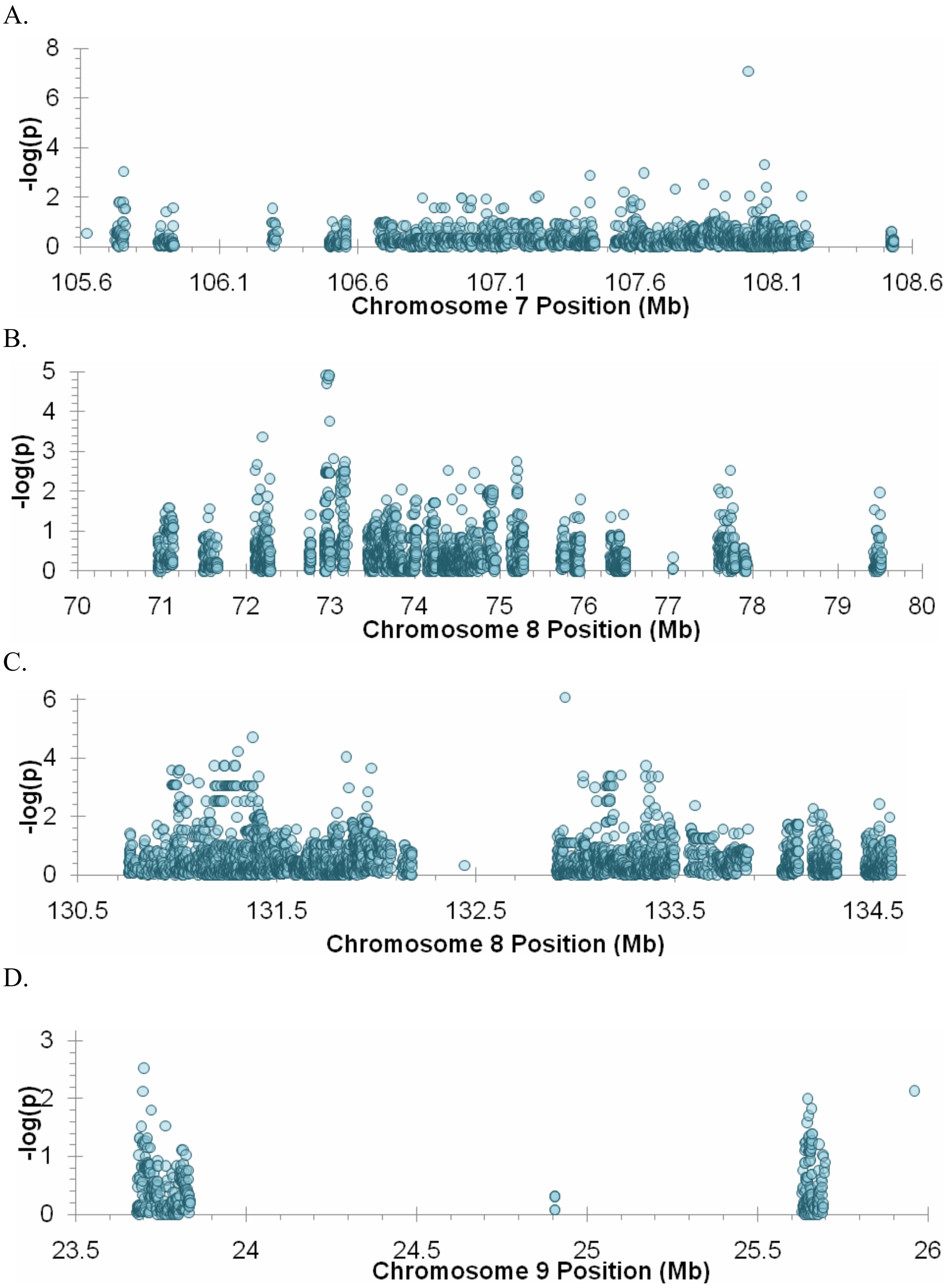


Figure 28. Allelic association study using 10 sequenced AP individuals as cases and CEU individuals sequenced for the 1000 genomes phase I project as controls. Results are shown for the (A) chromosome 7 (B) chromosome 8q21 (C) chromosome 8q24 and (D) chromosome 9 linkage regions. Physical positions are on hg19. P-values are not corrected for multiple testing.

Table 27. Top 50 most associated SNPs when using 10 AP as cases and 60 CEU individuals sequenced in the 1000 genomes pilot project as controls. P-values are not corrected for multiple testing. If SNP was not in dbSNP build 131, no rsName is listed. Positions are on hg19.

		SNP info				CEU			AP		Statistics	
Chr	Bp	rsName	Gene	Ref	Alt	#ref	#tot	depth	#ref	#tot	ChiSq	p
7	105753742	rs3757490	*SYPL1*	G	C	3	120	365	16	20	11.053	8.86E-04
7	108008678	rs12705466	*NRCAM*	A	C	8	120	96	10	20	28.731	8.32E-08
7	108067952	rs12705468	*NRCAM*	A	G	120	120	282	2	20	12.174	4.85E-04
8	72192850	.	*EYA1*	C	T	1	120	320	17	20	12.396	4.30E-04
8	72938357	rs2305018	*TRPA1*	A	C	16	120	339	9	20	19.119	1.23E-05
8	72953158	rs1025926	*TRPA1*	C	T	16	120	317	8	18	18.248	1.94E-05
8	72969263	rs3779752	*TRPA1*	A	C	8	120	216	12	20	18.817	1.44E-05
8	72970576	rs3779753	*TRPA1*	T	C	16	120	204	9	20	19.119	1.23E-05
8	72976997	rs16937961	*TRPA1*	C	T	16	120	347	9	20	19.119	1.23E-05
8	72987384	rs2278653	*TRPA1*	A	T	17	120	302	10	20	14.141	1.70E-04
8	130970406	.	*FAM49B*	G	A	4	120	240	15	20	13.379	2.55E-04
8	130974065	rs10111220	*FAM49B*	T	C	5	120	349	15	20	11.218	8.10E-04
8	130976675	.	*FAM49B*	G	A	5	120	325	15	20	11.218	8.10E-04
8	130976676	.	*FAM49B*	G	A	5	120	323	15	20	11.218	8.10E-04
8	130982439	rs16904182	*FAM49B*	C	T	5	120	322	15	20	11.218	8.10E-04
8	130983065	.	*FAM49B*	A	G	5	120	325	15	20	11.218	8.10E-04
8	130983869	rs16904183	*FAM49B*	A	G	5	120	269	15	20	11.218	8.10E-04
8	130985233	rs28682439	*FAM49B*	A	G	5	120	315	15	20	11.218	8.10E-04
8	130985309	rs28366859	*FAM49B*	A	G	5	120	355	15	20	11.218	8.10E-04
8	130990613	rs17194571	*FAM49B*	C	T	5	120	335	15	20	11.218	8.10E-04
8	130990662	rs17279655	*FAM49B*	T	C	5	120	314	15	20	11.218	8.10E-04
8	130996692	rs17194770	*FAM49B*	C	T	5	120	295	15	20	11.218	8.10E-04
8	130997011	.	*FAM49B*	G	T	5	120	265	15	20	11.218	8.10E-04
8	131002359	.	*FAM49B*	G	A	5	120	303	13	18	12.982	3.14E-04
8	131008342	.	*FAM49B*	G	A	4	120	355	15	20	13.379	2.55E-04
8	131011520	.	*FAM49B*	T	G	4	120	211	15	20	13.379	2.55E-04
8	131056570	rs11775966	*ASAP1*	C	T	8	120	199	9	14	12.076	5.11E-04
8	131105936	.	*ASAP1*	G	A	13	120	181	12	20	11.438	7.20E-04
8	131186515	.	*ASAP1*	A	T	2	120	366	16	20	14.046	1.78E-04
8	131236073	.	*ASAP1*	A	T	2	120	392	16	20	14.046	1.78E-04
8	131238322	.	*ASAP1*	C	T	2	120	248	16	20	14.046	1.78E-04
8	131283678	.	*ASAP1*	T	C	2	120	348	16	20	14.046	1.78E-04
8	131291187	rs11778881	*ASAP1*	T	C	2	120	384	16	20	14.046	1.78E-04
8	131304386	.	*ASAP1*	G	C	3	120	308	15	20	16.108	5.98E-05
8	131378870	.	*ASAP1*	G	T	1	120	369	16	20	18.286	1.90E-05
8	131407389	.	*ASAP1*	G	A	1	120	355	17	20	12.396	4.30E-04
8	131850824	rs57086962	*ADCY8*	G	T	4	120	238	13	18	15.341	8.97E-05
8	131975377	.	*ADCY8*	C	T	6	120	190	14	20	13.672	2.18E-04
8	132950043	.	*EFR3A*	C	T	10	120	201	10	20	24.306	8.22E-07
8	133041177	rs6471024	*OC90*	G	A	59	120	198	2	20	11.549	6.78E-04
8	133041179	rs6471025	*OC90*	A	G	57	120	193	2	20	12.449	4.18E-04
8	133168461	.	*KCNQ3*	C	T	1	120	314	17	20	12.396	4.30E-04
8	133175875	.	*KCNQ3*	G	A	1	120	281	17	20	12.396	4.30E-04
8	133183324	.	*KCNQ3*	C	T	1	120	305	17	20	12.396	4.30E-04
8	133229978	.	*KCNQ3*	C	A	12	120	242	12	20	12.600	3.86E-04
8	133353365	.	*KCNQ3*	G	A	1	120	305	17	20	12.396	4.30E-04
8	133357111	rs16904664	*KCNQ3*	G	C	2	120	247	16	20	14.046	1.78E-04
8	133382549	.	*KCNQ3*	C	A	1	120	333	17	20	12.396	4.30E-04
8	133419364	.	*KCNQ3*	C	A	1	120	253	17	20	12.396	4.30E-04
9	25643089	.	*TUSC1*	C	T	1	120	212	17	20	12.396	4.30E-04

Table 28. Ten novel non-synonymous SNPs detected.

Chr	Bp	Ref	Alt	AA change	Gene	SNPqual	Read Depth
7	107431599	G	A	Thr->Ile	SLC26A3	228	81
8	71581322	A	G	Ser->Pro	LACTB2	196	45
8	72229861	G	C	Thr->Arg	EYA1	185	99
8	73849653	A	G	His->Arg	KCNB2	228	86
8	75929334	G	A	Val->Ile	CRISPLD1	228	68
8	77764484	A	G	Glu->Gly	ZFHX4	228	66
8	133764112	T	C	Asp->Gly	TMEM71	228	89
8	134232908	C	T	Thr->Met	WISP1	228	35
8	134251224	C	T	Arg->His	NDRG1	179	33
8	134478308	T	C	Asn->Ser	ST3GAL1	228	59

Table 29. Novel SNPs detected in at least 3 of 10 sequenced AP individuals.

Chr	Bp	#	Ref	Alt	Gene
7	105733094	4	G	C	*SYPL1*
7	106812788	3	A	T	*HBP1*
7	106825004	3	C	G	*HBP1*
7	106969198	3	G	A	*COG5*
7	107036263	3	G	C	*COG5*
7	107847101	3	G	A	*NRCAM*
7	108150767	4	C	A	*PNPLA8*
8	73534238	3	T	G	*KCNB2*
8	73657598	5	A	C	*KCNB2*
8	74178643	3	T	C	*RPL7*
8	74435455	3	G	A	*STAU2*
8	74497692	3	A	G	*STAU2*
8	74658583	3	C	A	*STAU2*
8	74734062	3	C	A	*UBE2W*
8	74850445	3	C	T	*TCEB1*
8	75931846	3	G	A	*CRISPLD1*
8	76376046	3	G	A	*HNF4G*
8	76376262	3	A	G	*HNF4G*
8	76376289	3	G	A	*HNF4G*
8	77696239	3	T	C	*ZFHX4*
8	77769545	3	A	G	*ZFHX4*
8	130987057	4	C	T	*FAM49B*
8	130987422	4	A	G	*FAM49B*
8	130995940	4	A	C	*FAM49B*
8	131011075	5	A	G	*FAM49B*
8	131013835	4	C	T	*FAM49B*
8	131013846	3	T	C	*FAM49B*
8	131017116	5	G	A	*FAM49B*
8	131023523	4	T	C	*FAM49B*
8	131036353	4	G	A	*FAM49B*
8	131041020	4	C	T	*FAM49B*
8	131043310	3	T	C	*FAM49B*
8	131119997	4	A	G	*ASAP1*
8	131250556	4	G	C	*ASAP1*
8	131409449	4	C	T	*ASAP1*
8	131409630	3	G	A	*ASAP1*
8	131412805	3	G	A	*ASAP1*
8	131589997	3	C	T	*ASAP1*
8	131590060	3	G	A	*ASAP1*
8	131613767	3	T	G	*ADCY8*
8	131914317	3	A	C	*ADCY8*
8	131965372	3	G	T	*ADCY8*
8	132028015	4	G	A	*ADCY8*
8	132146925	3	A	G	*ADCY8*
8	132915056	3	G	A	*EFR3A*
8	133110672	3	A	G	*HHLA1*
8	133157906	4	T	C	*KCNQ3*
8	133160557	4	A	T	*KCNQ3*
8	133183324	3	C	T	*KCNQ3*
8	133194241	4	G	A	*KCNQ3*
9	23696856	3	T	G	*ELAVL2*
9	23716280	3	T	C	*ELAVL2*
9	23774019	3	C	T	*ELAVL2*

Table 30. Largest deletions detected in sequencing data. Basepair limits are on hg19. The EYA1 deletion has a frequency of 14.72% in the Database of Genomic Variants.[109]

Chr	Lower Limit	Upper Limit	Approx Size	# ind	# reads	Gene	Deletion in dbSNP or DGV nearby?
8	73787704	73793816	6112	10	326	KCNB2	rs66472029, 6058 bp
8	72214685	72217804	3119	4	47	EYA1	Variation_65163, 2963 bp
8	131850680	131852729	2049	6	112	ADCY8	rs6150816, 1988 bp
7	105733924	105734323	399	8	246	SYPL1	rs6150268, 338 bp
8	132168227	132168596	369	8	266	*ADCY8*	rs34226062, 313 bp
7	107683969	107684122	153	2	54	LAMB4	rs6150280, 143bp

Table 31. Case-control allelic association results for 27 promising SNPs in up to 31 AP singletons and up to 24 musically-trained individuals without AP of European descent. The SNP highlighted in red is a novel non-synonymous SNP, the 13 SNPs in purple are novel SNPs that were seen in at least 3 of the 10 originally sequenced individuals, and the 13 SNPs in yellow showed promising association signals in our initial association studies using 10 sequenced AP cases versus CEU controls genotyped for the HapMap or 1000 Genomes projects. Genomic positions are on hg19.

				Freq	#	Freq	# No				
Chr	Bp	SNP	A1	AP	AP	NoAP	AP	A2	ChiSq	P	OR
7	106812788	SNP106812788	T	0.19	31	0.353	20	A	3.048	0.0808	0.429
7	107592198	rs2072209	G	0.065	31	0.063	22	A	0.002	0.9657	1.034
7	107847101	SNP107847101	A	0.016	28	0	19	G	0.781	0.3767	NA
8	72969263	rs3779752	C	0.145	29	0.125	18	A	0.093	0.7599	1.189
8	72987384	rs2278653	A	0.29	31	0.196	22	T	1.262	0.2612	1.682
8	74178643	SNP74178643	C	0.048	29	0	17	T	2.388	0.1223	NA
8	74734062	SNP74734062	A	0.048	31	0.125	24	C	2.114	0.146	0.356
8	77696239	SNP77696239	C	0.032	31	0.063	24	T	0.57	0.4501	0.5
8	131008342	rs72720337	A	0	31	0.021	24	G	1.304	0.2536	0
8	131189298	rs11777289	C	0.032	31	0.083	23	T	1.369	0.2421	0.367
8	131250556	SNP131250556	C	0.032	31	0.068	24	G	0.739	0.39	0.456
8	131304386	rs77162821	C	0.032	31	0.083	24	G	1.369	0.2421	0.367
8	131378870	rs72724460	T	0.032	31	0.083	24	G	1.369	0.2421	0.367
8	131430133	rs7386870	T	0.468	31	0.304	24	C	2.942	0.0863	2.009
8	131613767	SNP131613767	G	0.226	31	0.229	24	T	0.002	0.9667	0.981
8	131850824	rs57086962	T	0.096	31	0.114	22	G	0.078	0.7799	0.83
8	131885942	rs12155610	T	0.371	31	0.548	24	C	3.168	0.0751	0.487
8	131965372	SNP131965372	T	0.067	31	0.091	24	G	0.21	0.6467	0.714
8	132028015	SNP132028015	A	0.048	31	0.063	23	G	0.105	0.7465	0.763
8	132146925	SNP132146925	G	0.032	31	0.063	24	A	0.57	0.4501	0.5
8	133157906	SNP133157906	C	0.133	26	0.175	22	T	0.327	0.5676	0.725
8	133183324	SNP133183324	T	0.113	31	0.159	21	C	0.479	0.4889	0.673
8	133357111	rs16904664	C	0.081	30	0.043	22	G	0.602	0.4379	1.93
8	133377440	rs2100646	A	0.207	31	0.184	24	G	0.074	0.785	1.155
8	133466893	rs3857927	T	0	31	0.045	24	G	2.872	0.0901	0
8	134478308	SNP134478308	C	0.016	30	0.021	20	T	0.034	0.8547	0.771
9	23696856	SNP23696856	G	0.097	31	0.045	22	T	0.971	0.3243	2.25

Table 32. Family-based allelic association results using genotype data from 121 AP family members (112 with AP, 5 without AP, and 4 with uncertain AP status) in 48 families, 31 unrelated AP cases, and 24 musically-trained non-AP controls of European descent. None of these uncorrected p-values would be significant after multiple testing correction. Results were obtained using the DFAM option in Plink.[82] Genomic coordinates are on hg19.

Chr	Bp	SNP	A1	A2	Obs	Exp	ChiSq	P
7	106812788	SNP106812788	T	A	21	23	0.8311	0.362
7	107592198	rs2072209	G	A	4	3.945	0.001829	0.9659
7	107847101	SNP107847101	A	G	1	0.5636	0.7742	0.3789
8	72969263	rs3779752	C	A	9	9.121	0.004274	0.9479
8	72987384	rs2278653	A	T	23	21.17	0.5876	0.4433
8	74178643	SNP74178643	C	T	3	1.691	2.366	0.124
8	74734062	SNP74734062	A	C	3	5.073	2.095	0.1478
8	77696239	SNP77696239	C	T	2	2.818	0.5651	0.4522
8	131008342	rs72720337	A	G	0	1.23	3.233	0.07216
8	131189298	rs11777289	C	T	6	6.382	0.07641	0.7822
8	131250556	SNP131250556	C	G	6	5.925	0.003415	0.9534
8	131304386	rs77162821	C	G	6	6.382	0.07641	0.7822
8	131378870	rs72724460	T	G	6	6.382	0.07641	0.7822
8	131430133	rs7386870	T	C	36	32.02	2.247	0.1338
8	131613767	SNP131613767	G	T	18	18.76	0.1144	0.7352
8	131850824	rs57086962	T	G	6	6.417	0.06319	0.8015
8	131885942	rs12155610	T	C	29	32.92	2.203	0.1377
8	131965372	SNP131965372	T	G	6	6.615	0.2081	0.6483
8	132028015	SNP132028015	A	G	3	3.382	0.1035	0.7476
8	132146925	SNP132146925	G	A	4	4.152	0.01632	0.8984
8	133110672	SNP133110672	G	A	2	1	2	0.1573
8	133157906	SNP133157906	C	T	9	10.33	0.5366	0.4639
8	133183324	SNP133183324	T	C	8	9.522	0.7238	0.3949
8	133357111	rs16904664	C	G	5	4.019	0.5962	0.44
8	133377440	rs2100646	A	G	15	15.65	0.0896	0.7647
8	133466893	rs3857927	T	G	0	2.17	4.799	0.02847
8	134478308	SNP134478308	C	T	1	1.127	0.03324	0.8553
9	23696856	SNP23696856	G	T	6	4.679	0.9622	0.3266

IX. MUSICAL TRAINING AND PITCH-NAMING ABILITY

Though much of our work on absolute pitch centered on the genetic components of its etiology, we also were interested in environmental influences on the trait, particularly musical training. It has been repeatedly demonstrated in the literature (reviewed in [46]) that absolute pitch possession is correlated with early musical training onset. However, these studies all had relatively small sample sizes and suggested slightly different age cutoffs before which musical training was initiated in the majority of AP possessors. With the online survey and pitch-naming test data gathered from a large number of participants (7,399) from February 2008 through March 2010, we sought to investigate the association of absolute pitch with early musical training in this population.

We also were interested in the accuracy and precision of pitch-naming by AP possessors and non-possessors and how this changes with age. Some investigation of the systematic shift in the sharp direction of pitch-naming by AP possessors with age had been undertaken previously[9] to validate anecdotal accounts of this phenomenon.[10] However, that study did not address the precision of AP participant pitch-naming nor the accuracy or precision of pitch-naming in individuals without AP.

Pitch-naming test scores and AP classification

The score distributions of our 7,399 study participants (Figure 29) resembled those of an earlier study of 2,213 participants using the same test and a shorter version of the survey.[9] The largest clusters of test scores occurred where one would expect to score by random chance (around 7.125) and above our threshold for AP possession, a pure tone score of 24.5. We classified each participant as an AP possessor (pure tone score ≥ 24.5), an individual with uncertain AP status (pure tone score >15 and <24.5), or a non-AP possessor (pure tone score ≤ 15). By these definitions, there were 2865 AP possessors, 943 participants of uncertain AP status, and 3591 non-AP possessors in our study who completed the survey and test (Figure 30A).

AP and musical training onset

When participants were divided into two subgroups, those who reported formal musical training and those who did not, it was evident that the majority of AP possessors received formal musical training (Figures 30B and 30C). It is likely that those individuals who reported no formal musical training but tested with AP learned the musical note names in a non-formal setting, which allowed them to demonstrate their pitch-naming abilities on our test. Early musical training onset was correlated with AP possession in our study (Figure 31). Though there was no absolute age cutoff before which musical training needed to be started in order for AP to develop, the majority of participants with AP received musical training before the age of 7 (Figures 31 and 32).

Pitch-naming accuracy and precision

Rather than classifying individuals as AP possessors by their test scores alone, we sought to investigate whether AP possessors had more accurate and/or precise test responses than did individuals with uncertain AP status and non-AP possessors. Each test response had a positive, negative, or zero deviation from the correct tone, in increments of 0.5 for each semitone. (Adjacent keys in Figure 1 are a semitone apart.) For instance, if the correct response was C and the participant's response was D#, the deviation would be 1.5, while if the participant's response was B it would be -0.5. The tritone (three full tones away from the correct response) was given a deviation value of 3.

Excluding notes for which there was no response, we calculated the mean deviation of each participant's test responses from the correct tones on the pure tone test. As expected, in non-AP possessors these mean deviations formed a continuous distribution centered on 0.25, the average deviation expected from random guessing. In contrast, AP possessors had a much tighter distribution centered closer to 0, with some skew in the positive direction (Figure 33). If instead the absolute value of each deviation was taken before averaging, we could determine the average deviation magnitude from the correct tone for each participant. The average deviation magnitude for AP possessors was 0.18 while the average deviation magnitude for non-AP possessors was 1.38, which is close to the average deviation magnitude of 1.5 expected by chance (Figure 34).

We were also interested in the precision of participant responses, so we next determined the spread of each participant's response deviation distribution by calculating the standard deviation of the pure tone test deviations. The vast majority of AP possessors named pitches a consistent difference away from the correct note, with a standard deviation of response deviations less than or equal to 1. Non-AP possessors generally had standard deviations greater than 1, with most falling near a standard deviation of 1.75, the expected value for random guessing. Individuals of uncertain AP status had a range of standard deviations below and above 1 (Figure 35).

Participants who made systematic errors on the pitch-naming test would be expected to have good precision like AP possessors but poor accuracy like non-AP possessors. One such systematic error we wanted to investigate was the shift in naming pitches too sharp with age.[9,10] As mentioned previously, in scoring we currently give 0.75 points for semitone errors to participants who are under 45, but we give full credit to individuals who are at least 45 if they make semi-tone errors. However, we have observed older participants make guesses that are consistently a tone or even two tones sharper than the correct response. As our test is currently designed, they would get no credit for those guesses. However, the standard deviation of their deviations from the correct responses would be low due to their consistency. Thus, we sought to determine if there was a greater trend towards sharper pitch-naming with age in individuals with low standard deviations of deviations as compared to those with higher standard deviations.

For this analysis, we first removed individuals who had listed their age as under 5 or over 100 or who answered less than 10 of the 36 scored tones on the pure tone test. As expected, the slope of the linear regression when mean deviations were plotted versus participant age was significantly different (F = 19.4759, DFn=1, DFd=7020, P<0.0001) in individuals with standard deviations of deviations that were less than or equal to 1 as compared to those with standard deviations greater than one (Figure 36). In the future, we could use the mean deviation and standard deviation of deviations to determine whether older participants have AP when they fail the initial test because of shifts greater than a semitone in their pitch-naming abilities with age or because of semitone shifts in participants under the age of 45.

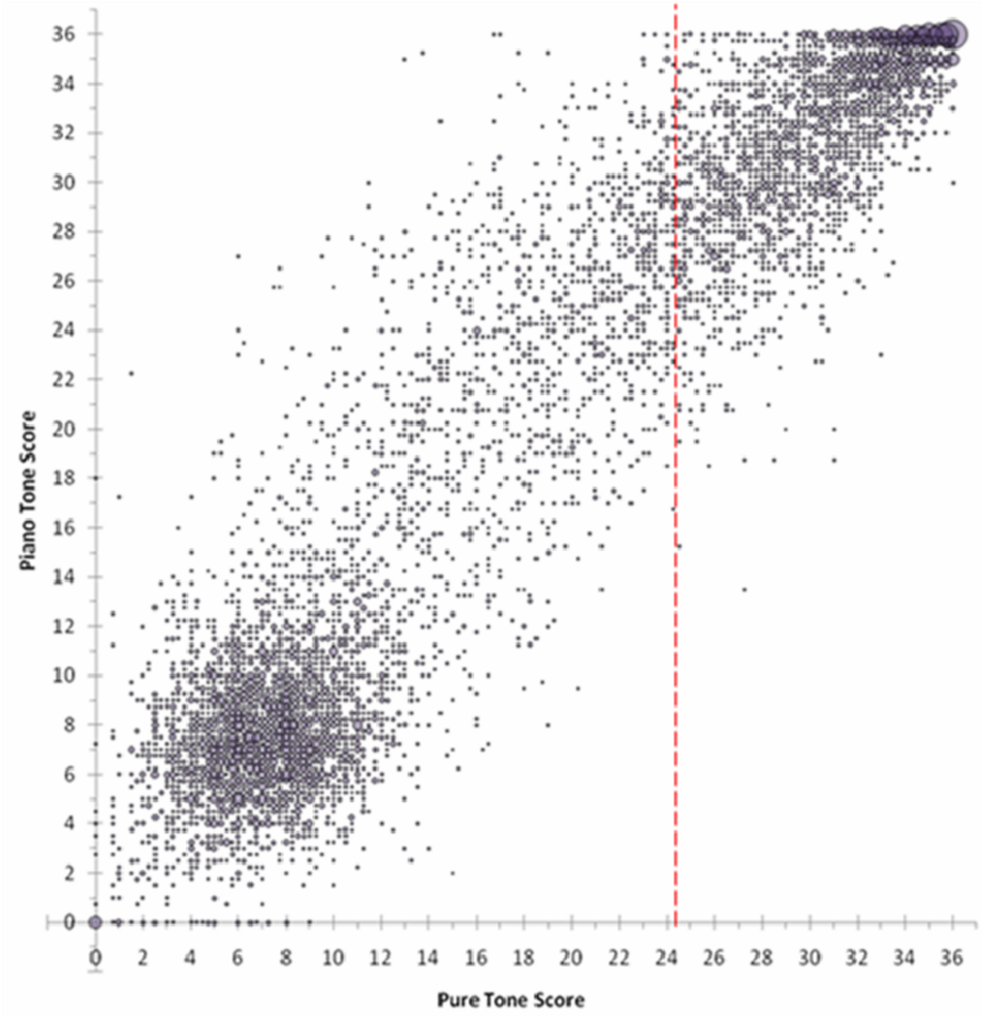


Figure 29. Piano tone test score versus pure tone test score for all participants. The area of the bubble is proportional to the number of participants who obtained that score combination. The red line indicates the threshold to be classified with AP in our study, which is a pure tone score of 24.5. The distribution is quite bimodal, with most participants scoring with AP or scoring near 7.125, the expected score with random guessing.

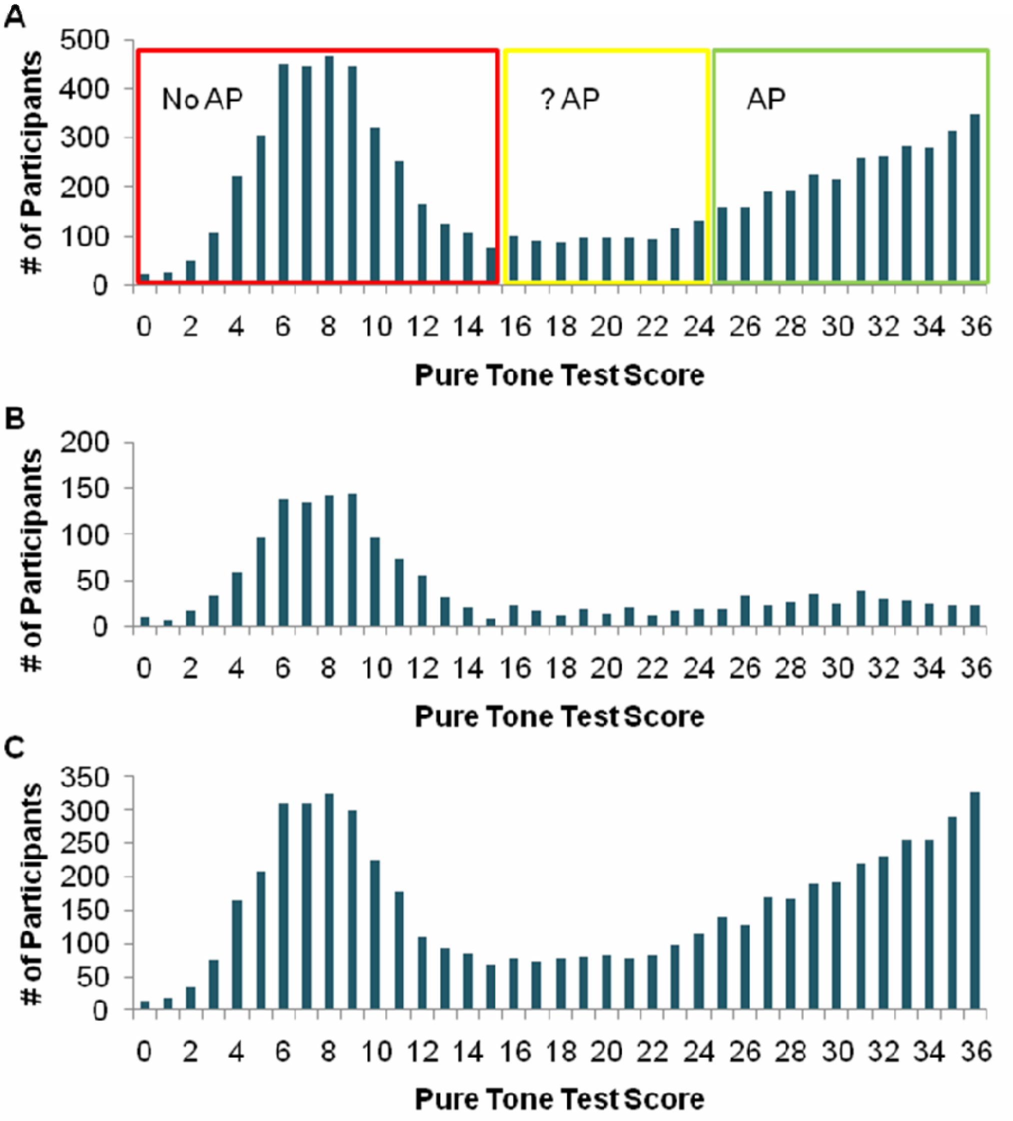


Figure 30. Pure tone test score distributions for (A) all participants (B) only participants who reported no formal musical training and (C) only participants who reported formal musical training. The cutoff for AP possession was a pure tone test score of 24.5.

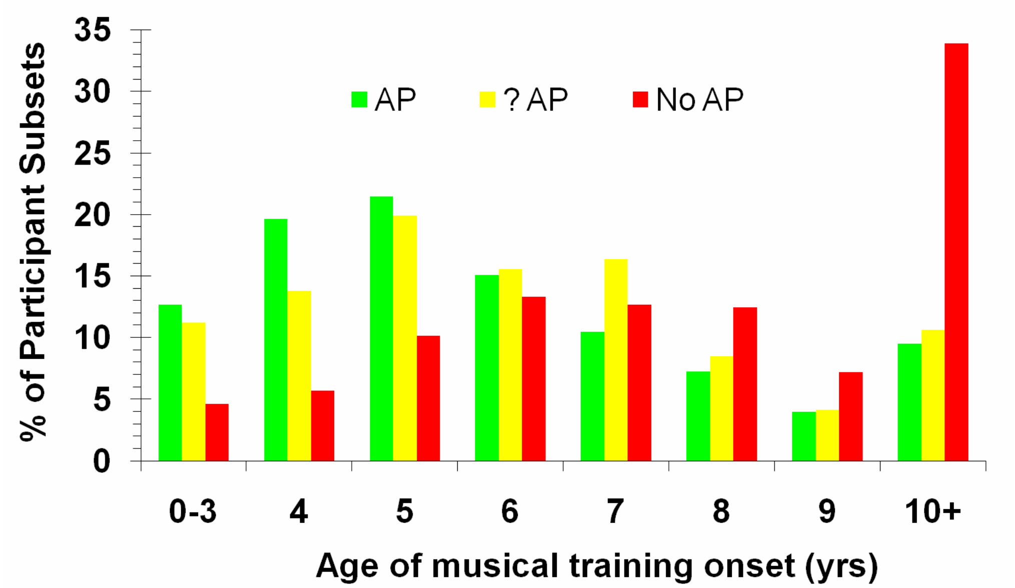


Figure 31. Proportions of participants who tested with AP, had an unknown AP status, or tested without AP divided by the age at which they began formal musical training.

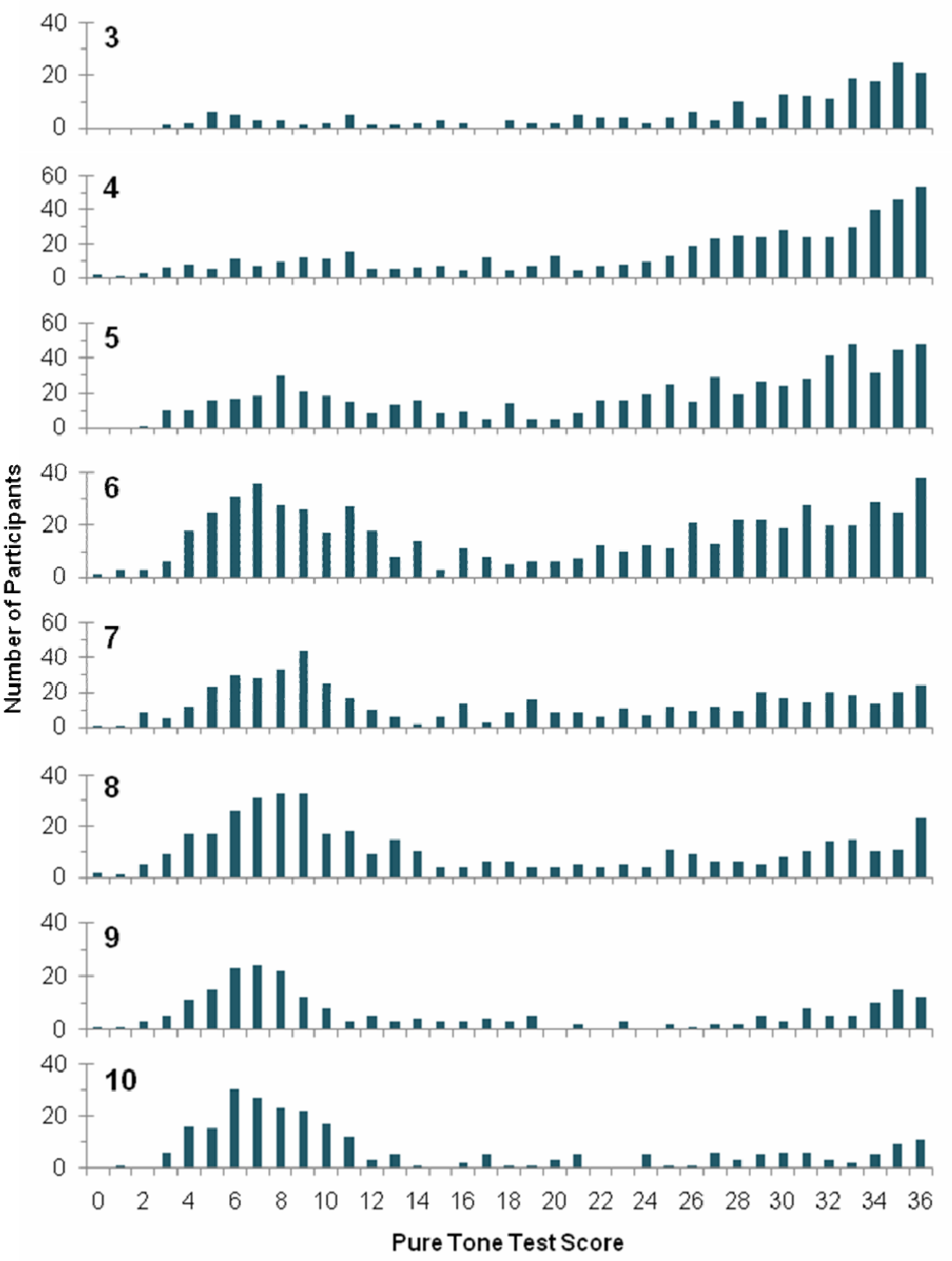


Figure 32. Distributions of pure tone test scores divided by age (3-10) of participant when they began formal musical training.

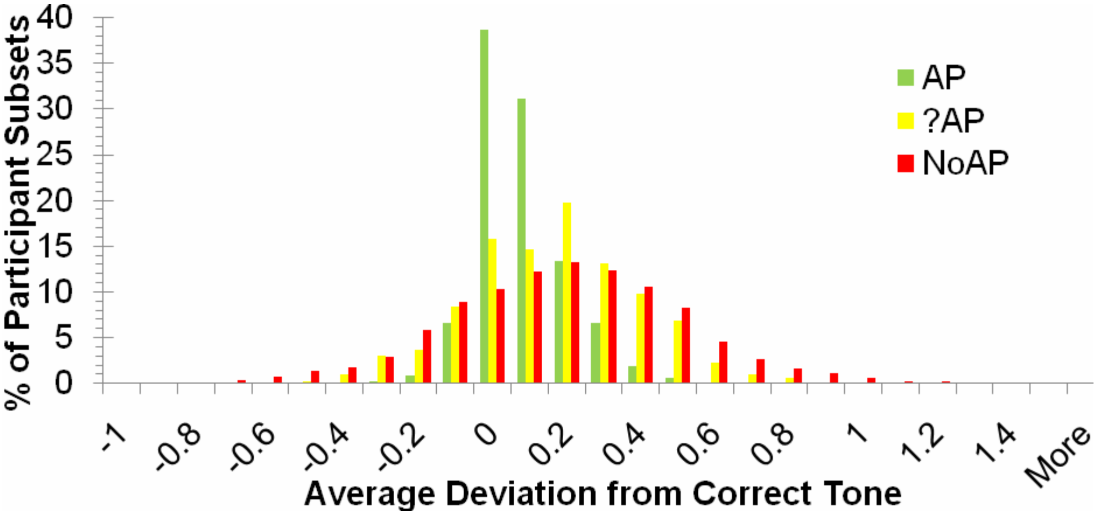


Figure 33. Average deviations of participant responses from the correct tones for participants with AP, with unknown AP status, or without AP. An average deviation of 0.25 is expected by chance.

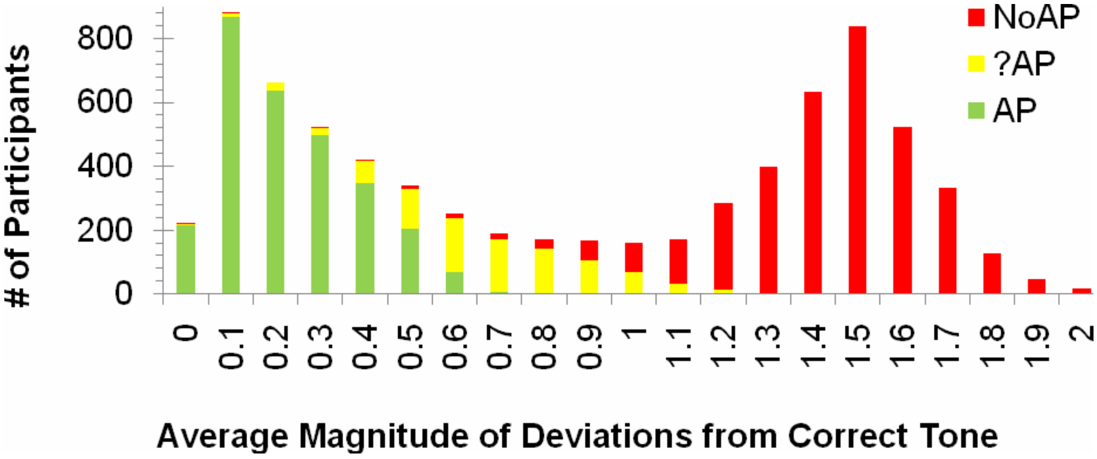


Figure 34. Average of the absolute values of deviations of participant responses from the correct tones for participants with AP, with unknown AP status, or without AP. An average deviation magnitude of 1.5 is expected by chance.

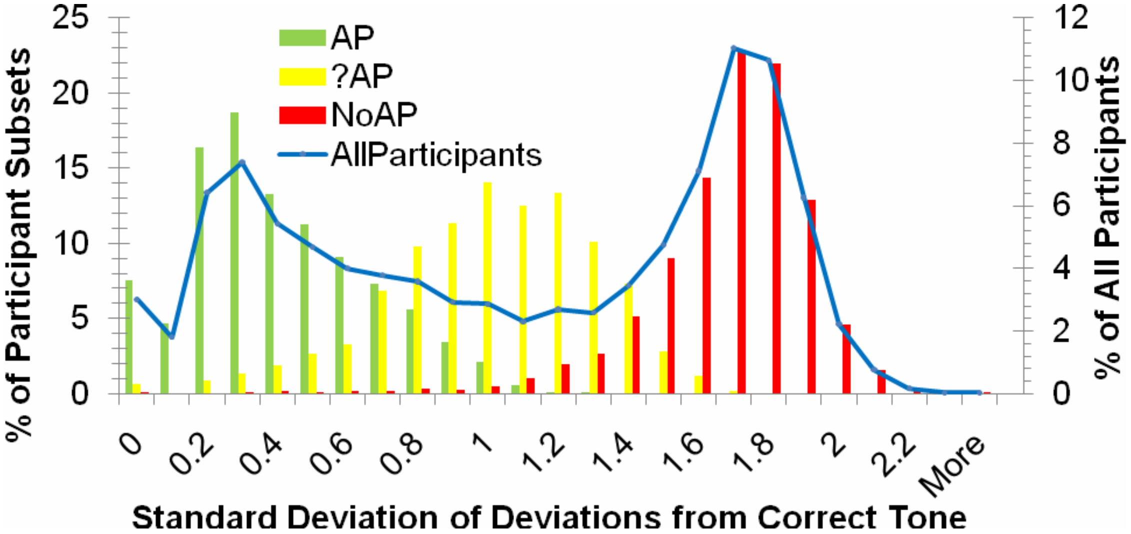


Figure 35. Standard deviation of deviations of participant responses from the correct tones for participants with AP, with unknown AP status, or without AP. A standard deviation of 1.75 is expected by chance.

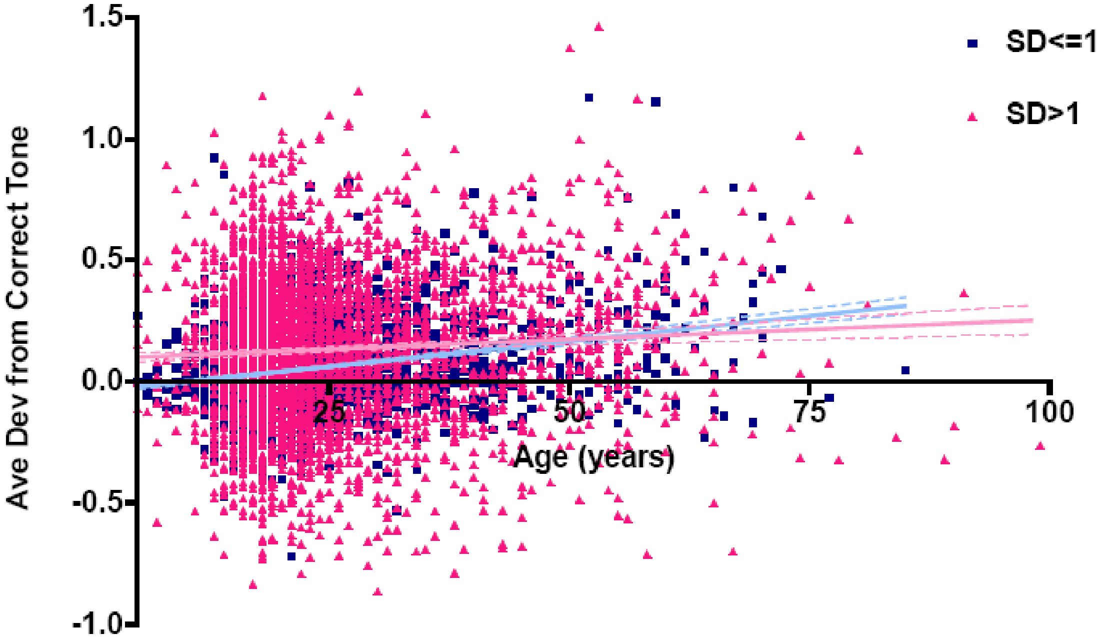


Figure 36. Sharpward shift in pitch-naming with age. The linear regression equation for individuals with standard deviations of deviations from the correct tones that were less than or equal to 1 was y = 0.004173245x - 0.04100061, while the line for individuals with standard deviations greater than 1 was y = 0.001656785 x + 0.09260276. The difference in the slopes of the two lines was extremely significant (p < 0.0001), and both slopes were also significantly different than 0 (p < 0.0001). Dashed lines show 95% confidence intervals of the linear regressions.

X. CONCLUSIONS

Absolute pitch is a complex trait, and pinpointing genetic variants that predispose individuals to developing the trait proved to be more difficult than expected. Not only do environmental factors, such as early musical training, appear to impact the penetrance of AP, but the genetic factors involved are likely more numerous, more heterogeneous, and of lower effect size than we had originally hoped. Although one genetic variant in any of a variety of genes could predispose individuals to developing AP by itself, it is also possible that gene-gene interactions might be required for the development of AP, increasing the complexity even further.

One fruitful aspect of our study was the detection of one significant and three suggestive regions of linkage to AP in families of European descent (Chapter IV). Though there were relatively few genes near the significant linkage peak, we were unable to pinpoint genetic variants in the region that were convincingly associated with AP, either by candidate gene sequencing (Chapter VII) or targeted next-generation sequencing (Chapter VIII). Perhaps multiple genetic variants were contributing to each linkage peak, making them harder to find. The suggestive linkage region on 8q21.11 at 86.7 cM was close to a peak (chromosome 8 at 92 cM) reported in a recent linkage study on Finnish families with musical-aptitude,[110] so it is possible that some of the genetic factors that play a role in AP also play a role in other aspects of musical ability.

The intertwining of genetic factors that influence AP and music ability in general is one possible explanation for the puzzling observation that allele frequencies for some genetic variants differ greatly between the population of CEU individuals with unknown

AP status and an unknown degree of musical training used as controls and the smaller numbers of musically-trained controls without AP that we collected (Chapter VII). Perhaps factors that by themselves predispose to AP but are not sufficient to cause AP are enriched in individuals with musical training, because children are more likely to initiate and/or continue music lessons if they have some inborn musical affinity or ability.

Though our twin studies indicated that genetic factors play a role in AP, the majority of AP possessors who participated in our study were singletons without any relatives with AP. Allele frequencies in AP singletons versus AP probands from multiplex families also differed at some loci (Chapter VII), indicating that different genetic factors may be at work in the families or that genetic factors may play a proportionally larger role than environmental factors in multiplex AP families than AP singletons.

The lack of fruitfulness of some aspects of our genetic study could have been due to sample size or financial and time limitations. In our linkage studies (Chapters III-V), we were able to include all of the informative multiplex families that we were able to recruit over the past ten years, but they were still not powered adequately, and we did not encounter any very large families for analysis. The Ashkenazi Jewish GWAS was extensively underpowered, even after the addition of control data from other studies, which subsequently introduced additional chances for errors (Chapter VI). We simply did not have the funds to recruit additional participants and subsequently genotype them. Finally, our next-generation sequencing study (Chapter VIII) produced many more leads than we could follow up on here. We could potentially have refined our list of candidates by sequencing greater than ten AP individuals initially, perhaps including some controls or family members as well, but that would have been more expensive.

In the future, human genomes will be sequenced on a continuously larger scale, starting with the 1000 Genomes Project which is currently underway. This project and others will make whole genome sequence data from multitudes of individuals available publicly, which would allow us to reanalyze our own large-scale sequencing data by comparing it with allele frequencies of increasingly less common alleles. Thus, the potential to detect genetic variants that are associated with AP using our existing data still exists. In the future, additional individuals with AP, their family members, and/or musically-trained individuals without AP could be sequenced on a genome-wide level, if enough resources would be available to do so.

In the event that AP-predisposing genetic variants are discovered, their prevalence could be estimated in various human sub-populations to determine if they explain the differences in prevalence reported in different populations. Evolutionarily, determining if non-human primates and extensive vocal communicators, such as songbirds, dolphins, whales, and bats, have similar variants would indicate if these variants arose in the human population or if they are common to species that rely on sound perception for communication. Genetically, introducing the variants into a genetic model system, such as a mouse, could allow a more extensive characterization of the anatomical and functional effects of the variants on the brain.

In addition to elucidating more about the etiology of absolute pitch, this study also illustrated some important factors to consider when studying any trait genetically. Throughout our study, we realized that looking at raw data was crucial to eliminate false positives and false negatives. This was true of microsatellite genotype data, SNP genotype data, Sanger sequencing data, and next-generation sequencing data. Though software analysis tools continue to improve, sometimes there is no substitute for human inspection, and putting full trust in calls made by the computer could be the downfall of a study. This becomes increasingly important as data quantity blossoms, and it is tempting to believe the results the computer gives you without any follow-up.

Also, using a website as a recruiting tool allowed us to at least begin to study a rare, complex trait because we gathered a sufficient number of participants. Though families in which there were multiple AP possessors were limited, if we had more resources, we could have collected DNA samples from hundreds if not thousands of the singleton AP possessors in our database. Critics of using websites in this manner cite cheating on the test, lying on the survey, and lack of security to protect personal data as major obstacles, but I think that the sheer number of participants minimized the damage done by the first two factors and that we took reasonable security measures, given that we did not collect medical records or other sensitive data. It is true that a number of the participants in our survey and test were not willing to contribute a DNA sample to our study, but I think this was more due to the fact that AP is not a disease trait than the fact that the participants were recruited via a website.

As with other complex traits and diseases, there is still a lot to learn about the etiology of absolute pitch. Hopefully future studies will incorporate more participants and have enough power to look at the interplay of many factors in this trait. Though absolute pitch is a rare, binary trait, the complexities that we have encountered in studying its causes remind us that genetics, cognition, and the brain can be far from simple.

REFERENCES

1. Stumpf, C. (1883) Tonpsychologie. S. Hirzel, Leipzig.
2. Ward, W.D. (1999) Absolute Pitch. In: Deutsch D (ed) The Psychology of Music. Academic Press, pp 265-298.
3. Takeuchi, A.H., and Hulse, S.H. (1993). Absolute Pitch. Psychol. Bull. *113*, 345-361.
4. Terhardt, E., and Ward, W.D. (1982). Recognition of musical key: exploratory study. J. Acoust. Soc. Am. *72*, 26-33.
5. Levitin, D.J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. Percept. Psychophys. *56*, 414-423.
6. Bachem, A. (1937). Various types of absolute pitch. The Journal of the Acoustical Society of America *9*, 146-151.
7. Deutsch, D., Henthorn, T., and Dolson, M. (2004). Absolute pitch, speech, and tone language: some experiments and a proposed framework. Music Perception *21*, 339-356.
8. Baharloo, S., Johnston, P.A., Service, S.K., Gitschier, J., and Freimer, N.B. (1998). Absolute pitch: an approach for identification of genetic and nongenetic components. Am. J. Hum. Genet. *62*, 224-231.
9. Athos, E.A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., and Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. Proc. Natl. Acad. Sci. U. S. A. *104*, 14795-14800.
10. Vernon, P.E. (1977). Absolute pitch: a case study. Br. J. Psychol. *68*, 485-489.
11. Petran, L.A. (1932). An experimental study of pitch recognition. Psychol. Monogr. *42*, 1-120.
12. Hartman, E.B. (1954). The influence of practice and pitch-distance between tones on the absolute identification of pitch. Am. J. Psychol. *67*, 1-14.
13. Balzano, G.J. (1984). Absolute pitch and pure tone identification. J. Acoust. Soc. Am. *75*, 623-625.
14. Heaton, P. (2003). Pitch memory, labelling and disembedding in autism. J. Child Psychol. Psychiatry *44*, 543-551.
15. Ross, D.A., Olson, I.R., Marks, L.E., and Gore, J.C. (2004). A nonmusical paradigm for identifying absolute pitch possessors. J. Acoust. Soc. Am. *116*, 1793-1799.
16. Saffran, J.R., and Griepentrog, G.J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. Dev. Psychol. *37*, 74-85.
17. Weisman, R., Njegovan, M., Sturdy, C., Phillmore, L., Coyle, J., and Mewhort, D. (1998). Frequency-range discriminations: Special and general abilities in zebra finches (Taeniopygia guttata) and humans (Homo sapiens). J. Comp. Psychol. *112*, 244-258.
18. Weisman, R.G., Njegovan, M.G., Williams, M.T., Cohen, J.S., and Sturdy, C.B. (2004). A behavior analysis of absolute pitch: sex, experience, and species. Behavioural processes *66*, 289-307.
19. Chaloupka, V., Mitchell, S., and Muirhead, R. (1994). Observation of a reversible, medication-induced change in pitch perception. J. Acoust. Soc. Am. *96*, 145-149.

20. Kobayashi, T., Nisijima, K., Ehara, Y., Otsuka, K., and Kato, S. (2001). Pitch perception shift: a rare side-effect of carbamazepine. Psychiatry Clin. Neurosci. *55*, 415-417.

21. Konno, S., Yamazaki, E., Kudoh, M., Abe, T., and Tohgi, H. (2003). Half pitch lower sound perception caused by carbamazepine. Intern. Med. *42*, 880-883.

22. Koerner, C., and Deuschle, M. (2003). Reversible pitch perception shift caused by trimipramine. Pharmacopsychiartry *36*, 241.

23. Wynn, V.T. (1971). "Absolute" pitch--a bimensual rhythm. Nature *230*, 337.

24. Bhatt, K.A., Liberman, M.C., and Nadol, J.B., Jr. (2001). Morphometric analysis of age-related changes in the human basilar membrane. Ann. Otol. Rhinol. Laryngol. *110*, 1147-1153.

25. Bachem, A. (1955). Absolute pitch. The Journal of the Acoustical Society of America *27*, 1180-1185.

26. Hamilton, R.H., Pascual-Leone, A., and Schlaug, G. (2004). Absolute pitch in blind musicians. Neuroreport *15*, 803-806.

27. Gregersen, P.K., Kowalsky, E., Kohn, N., and Marvin, E.W. (1999). Absolute pitch: Prevalence, ethnic variation, and estimation of the genetic component. Am. J. Hum. Genet. *65*, 911-913.

28. Gregersen, P.K., Kowalsky, E., Kohn, N., and Marvin, E.W. (1999). Absolute pitch: prevalence, ethnic variation, and estimation of the genetic component. Am. J. Hum. Genet. *65*, 911-913.

29. Heaton, P., Hermelin, B., and Pring, L. (1998). Autism and pitch processing: A precursor for savant musical ability? Music Perception *15*, 291-305.

30. Lenhoff, H.M. (2001). Williams syndrome. Scientist *15*, 6-6.

31. Brown, W.A., Cammuso, K., Sachs, H., Winklosky, B., Mullane, J., Bernier, R., Svenson, S., Arin, D., Rosen-Sheidley, B., and Folstein, S.E. (2003). Autism-related language, personality, and cognition in people with absolute pitch: results of a preliminary study. J. Autism Dev. Disord. *33*, 163-167; discussion 169.

32. Griffiths, T.D. (2001). The neural processing of complex sounds. Ann. N. Y. Acad. Sci. *930*, 133-142.

33. Liegeois-Chauvel, C., Giraud, K., Badier, J.M., Marquis, P., and Chauvel, P. (2001). Intracerebral evoked potentials in pitch perception reveal a functional asymmetry of the human auditory cortex. Ann. N. Y. Acad. Sci. *930*, 117-132.

34. Liegeois-Chauvel, C., de Graaf, J.B., Laguitton, V., and Chauvel, P. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. Cereb. Cortex *9*, 484-496.

35. Zatorre, R.J. (2001). Neural specializations for tonal processing. Ann. N. Y. Acad. Sci. *930*, 193-210.

36. Tervaniemi, M., and Hugdahl, K. (2003). Lateralization of auditory-cortex functions. Brain Res. Brain Res. Rev. *43*, 231-246.

37. Schlaug, G., Jancke, L., Huang, Y.X., and Steinmetz, H. (1995). In-vivo evidence of structural brain asymmetry in musicians. Science *267*, 699-701.

38. Keenan, J.P., Thangaraj, V., Halpern, A.R., and Schlaug, G. (2001). Absolute pitch and planum temporale. Neuroimage *14*, 1402-1408.

39. Zatorre, R.J. (1989). Intact absolute pitch ability after left temporal lobectomy. Cortex *25*, 567-580.

40. Sutton, S., Braren, M., Zubin, J., and John, E.R. (1965). Evoked-potential correlates of stimulus uncertainty. Science *150*, 1187-1188.
41. Klein, M., Coles, M.G.H., and Donchin, E. (1984). People with absolute pitch process tones without producing a P300. Science *223*, 1306-1309.
42. Wayman, J.W., Frisina, R.D., Walton, J.P., Hantz, E.C., and Crummer, G.C. (1992). Effects of musical training and absolute pitch ability on event-related activity in response to sine tones. J. Acoust. Soc. Am. *91*, 3527-3531.
43. Zatorre, R.J., Perry, D.W., Beckett, C.A., Westbury, C.F., and Evans, A.C. (1998). Functional anatomy of musical processing in listeners with absolute pitch and relative pitch. Proc. Natl. Acad. Sci. U. S. A. *95*, 3172-3177.
44. Sergeant, D. (1969). Experimental investigation of absolute pitch. J. Res. Music Educ. *17*, 135-143.
45. Miyazaki, K. (1988). Musical pitch identification by absolute pitch possessors. Percept. Psychophys. *44*, 501-512.
46. Takeuchi, A.H., and Hulse, S.H. (1993). Absolute pitch. Psychol. Bull. *113*, 345-361.
47. Gregersen, P.K., Kowalsky, E., Kohn, N., and Marvin, E.W. (2001). Early childhood music education and predisposition to absolute pitch: teasing apart genes and environment. Am. J. Med. Genet. *98*, 280-282.
48. Deutsch, D., Dooley, K., Henthorn, T., and Head, B. (2009). Absolute pitch among students in an American music conservatory: association with tone language fluency. J. Acoust. Soc. Am. *125*, 2398-2403.
49. Bachem, A. (1940). The genesis of absolute pitch. J. Acoust. Soc. Am. *11*, 434-439.
50. Profita, J., and Bidder, T.G. (1988). Perfect pitch. Am. J. Med. Genet. *29*, 763-771.
51. Baharloo, S., Service, S.K., Risch, N., Gitschier, J., and Freimer, N.B. (2000). Familial aggregation of absolute pitch. Am. J. Hum. Genet. *67*, 755-758.
52. Gregersen, P.K. (1998). Instant recognition: The genetics of pitch perception. Am. J. Hum. Genet. *62*, 221-223.
53. Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314-331.
54. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L., and Weber, J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am. J. Hum. Genet. *63*, 861-869.
55. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304-1351.
56. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
57. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241-247.

58. Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. *46*, 229-241.

59. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

60. Witte, J.S., Carlin, J.B., and Hopper, J.L. (1999). Likelihood-based approach to estimating twin concordance for dichotomous traits. Genet. Epidemiol. *16*, 290-304.

61. McGue, M. (1992). When assessing twin concordance, use the probandwise not the pairwise rate. Schizophr. Bull. *18*, 171-176.

62. Drayna, D., Manichaikul, A., de Lange, M., Snieder, H., and Spector, T. (2001). Genetic correlates of musical pitch recognition in humans. Science *291*, 1969-1972.

63. Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. U. S. A. *102*, 10604-10609.

64. Davie, A.M. (1979). The 'singles' method for segregation analysis under incomplete ascertainment. Ann. Hum. Genet. *42*, 507-512.

65. Stellingwerff, H.J., van Hagen, J.M., and ten Kate, L.P. (2006). Segregation ratio in cranio-cerebello-cardiac syndrome. Eur. J. Hum. Genet. *14*, 1054-1057.

66. Lenhoff, H.M., Wang, P.P., Greenberg, F., and Bellugi, U. (1997). Williams syndrome and the brain. Sci. Am. *277*, 68-73.

67. Lenhoff, H.M., Perales, O., and Hickok, G. (2001). Absolute pitch in Williams syndrome. Music Perception *18*, 491-503.

68. Gregersen, P.K., Kowalsky, E., de Andrade, M., and Jawaheer, D. (1997). Affected sib pair analysis of families with absolute pitch (AP); exclusion of the Williams locus. Am. J. Hum. Genet. *61*, A399-A399.

69. Freimer, N., and Sabatti, C. (2004). The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat. Genet. *36*, 1045-1051.

70. Greenberg, D.A., and Abreu, P.C. (2001). Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. Genet. Epidemiol. *21*, 299-314.

71. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. *30*, 97-101.

72. O'Connell, J.R., and Weeks, D.E. (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am. J. Hum. Genet. *63*, 259-266.

73. Nievergelt, C.M., Smith, D.W., Kohlenberg, J.B., and Schork, N.J. (2004). Large-scale integration of human genetic and physical maps. Genome Res. *14*, 1199-1205.

74. Lander, E.S., and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. U. S. A. *84*, 2363-2367.

75. Elston, R.C., and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. Hum. Hered. *21*, 523-542.

76. Whittemore, A.S., and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. Biometrics *50*, 118-127.

77. Kong, A., and Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. Am. J. Hum. Genet. *61*, 1179-1188.

78. Sullivan, P.F., Neale, B.M., Neale, M.C., van den Oord, E., and Kendler, K.S. (2003). Multipoint and single point non-parametric linkage analysis with imperfect data. Am J Med Genet B Neuropsychiatr Genet *121B*, 89-94.

79. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241-247.

80. Theusch, E., Basu, A., and Gitschier, J. (2009). Genome-wide study of families with absolute pitch reveals linkage to 8q24.21 and locus heterogeneity. Am. J. Hum. Genet. *85*, 112-119.

81. Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. Hum. Genet. *73*, 320-326.

82. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559-575.

83. Abecasis, G.R., and Wigginton, J.E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am. J. Hum. Genet. *77*, 754-767.

84. Boyles, A.L., Scott, W.K., Martin, E.R., Schmidt, S., Li, Y.J., Ashley-Koch, A., Bass, M.P., Schmidt, M., Pericak-Vance, M.A., Speer, M.C., et al. (2005). Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. Hum. Hered. *59*, 220-227.

85. Camp, N.J., and Farnham, J.M. (2001). Correcting for multiple analyses in genomewide linkage studies. Ann. Hum. Genet. *65*, 577-582.

86. Brown, M.T., Andrade, J., Radhakrishna, H., Donaldson, J.G., Cooper, J.A., and Randazzo, P.A. (1998). ASAP1, a phospholipid-dependent arf GTPase-activating protein that associates with and is phosphorylated by Src. Mol. Cell. Biol. *18*, 7038-7051.

87. Ludwig, M.G., and Seuwen, K. (2002). Characterization of the human adenylyl cyclase gene family: cDNA, gene structure, and tissue distribution of the nine isoforms. J. Recept. Signal Transduct. Res. *22*, 79-110.

88. Wong, S.T., Athos, J., Figueroa, X.A., Pineda, V.V., Schaefer, M.L., Chavkin, C.C., Muglia, L.J., and Storm, D.R. (1999). Calcium-stimulated adenylyl cyclase activity is critical for hippocampus-dependent long-term memory and late phase LTP. Neuron *23*, 787-798.

89. de Quervain, D.J., and Papassotiropoulos, A. (2006). Identification of a genetic cluster influencing memory performance and hippocampal activity in humans. Proc. Natl. Acad. Sci. U. S. A. *103*, 4270-4274.

90. Risch, N. (1990). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am. J. Hum. Genet. *46*, 242-253.

91. Wiltshire, S., Cardon, L.R., and McCarthy, M.I. (2002). Evaluating the results of genomewide linkage scans of complex traits by locus counting. Am. J. Hum. Genet. *71*, 1175-1182.

92. Abecasis, G.R., Burt, R.A., Hall, D., Bochum, S., Doheny, K.F., Lundy, S.L., Torrington, M., Roos, J.L., Gogos, J.A., and Karayiorgou, M. (2004). Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. Am. J. Hum. Genet. *74*, 403-417.

93. Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., and Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat. Genet. *9*, 152-159.

94. Wright, A.F., Carothers, A.D., and Pirastu, M. (1999). Population choice in mapping genes for complex diseases. Nat. Genet. *23*, 397-404.

95. Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. *12*, 771-776.

96. Mitchell, M.K., Gregersen, P.K., Johnson, S., Parsons, R., and Vlahov, D. (2004). The New York Cancer Project: rationale, organization, design, and baseline characteristics. J. Urban Health *81*, 301-310.

97. Need, A.C., Kasperaviciute, D., Cirulli, E.T., and Goldstein, D.B. (2009). A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. Genome biology *10*, R7.

98. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

99. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U. S. A. *101*, 6062-6067.

100. Fletcher, G.C., Patel, S., Tyson, K., Adam, P.J., Schenker, M., Loader, J.A., Daviet, L., Legrain, P., Parekh, R., Harris, A.L., et al. (2003). hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene C4.4a and dystroglycan. Br. J. Cancer *88*, 579-585.

101. Veenstra-VanderWeele, J., Kim, S.J., Gonen, D., Hanna, G.L., Leventhal, B.L., and Cook, E.H., Jr. (2001). Genomic organization of the SLC1A1/EAAC1 gene and mutation screening in early-onset obsessive-compulsive disorder. Mol. Psychiatry *6*, 160-167.

102. Shan, Z., Parker, T., and Wiest, J.S. (2004). Identifying novel homozygous deletions by microsatellite analysis and characterization of tumor suppressor candidate 1 gene, TUSC1, on chromosome 9p in human lung cancer. Oncogene *23*, 6612-6620.

103. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid

selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182-189.

104. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

105. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

106. The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

107. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

108. The 1000 Genomes Project (2010) http://www.1000genomes.org.

109. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2009). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704-712.

110. Pulli, K., Karma, K., Norio, R., Sistonen, P., Goring, H.H., and Jarvela, I. (2008). Genome-wide linkage scan for loci of musical aptitude in Finnish families: evidence for a major locus at 4q22. J. Med. Genet. *45*, 451-456.

Appendix A. Updated absolute pitch survey put online (http://perfectpitch.ucsf.edu) in February 2008. *All of the information you provide on this survey will be stored in a secure, password-protected database and will be kept confidential. The data will only be used for our absolute pitch survey and participants will be anonymous in any results we may publish. Please contact us at perfect.pitch@ucsf.edu if you have any questions or concerns about this. By filling out this survey, you are providing your consent to participate in the University of California Genetics of Absolute Pitch Study.*

Contact Information – *This allows us to keep in touch with our participants and allows us to re-contact individuals who we would like to provide more information or participate further in our study.*
First Name: (required)
Last Name: (required)
Phone Number:
Email: (required)
Street Address:
City:
State:
ZIP:
Country of current residence:
Would you like to be notified of our study results in the future? Yes/No
Would you like to participate further in our study? Yes/No

Demographic Information – *This allows us to compare individuals within and across demographic groups and dissect the effects of language, country of origin, and ethnic background on the acquisition of absolute pitch.*
Age (required):
Gender: *Female / Male*
Which do you consider yourself? *Right-handed / Left-handed / Ambidextrous*
Are you a twin or multiple birth? *Yes / No*
If you are a twin or multiple birth, please specify which type: *Identical / Fraternal*
There may be correlations between language exposure and absolute pitch.
What is your native language (the first language you learned to speak)?
Are you multi-lingual? *Yes/No*
If so, please describe any other languages you have learned, and the age at which you started learning them.
Are you a tonal language speaker? *Yes/No/Don't Know*
The cultures of different countries may influence the likelihood that individuals will develop absolute pitch.
Which country were you born in?
Did you live in any other countries while growing up? *Yes/No*
If so, please list any additional countries and approximately how old you were when you lived there:
Ethnicities can differ both genetically and culturally, and we are interested in the effect of ethnicity on the acquisition of absolute pitch.
Please indicate the ethnicity(ies) that best describe you (examples include Polish, African American, Korean, Ashkenazi Jewish):
If you would prefer not to disclose this information, please check here → ___

Please list the ethnicities/countries of origin of each of your 4 grandparents to the best of your ability:
Maternal grandmother: *Mat. grandfather:*
Paternal grandmother: *Pat. grandfather:*

Musical Training History – *Musical training is an important ingredient for developing absolute pitch. We are interested to learn the extent of musical training of our participants and whether any particular type of training correlates with pitch naming ability.*
Have you had formal musical training (music lessons)? *Yes/No*
If so, how old were you when the musical training began?
Why did you begin musical training? *Self-motivated/Parent-motivated/Other* Explain:
On what instrument did you begin musical training?
Which instrument(s) do you currently play?
If you have had musical training before age 6, did it include ear training? *Yes / No / Don't Know*
How many years of training have you had? *Less than 1 / 1-5 / 6-10 / 11-20 / More than 20*
Have you studied the Suzuki method? *Yes / No*
Do you primarily assign English letters (A,B,C,etc.) or solfeggio designations (do,re,mi,etc.) to notes? *English/Solfeggio/Both/Neither*
If solfeggio, do you use Fixed-Do or Movable-Do designations? *Fixed/Movable/Don't Know*
If you are a twin or multiple birth, how much musical training did your multiple birth/twin sibling(s) have? *About the same as you/more than you/less than you* Please elaborate:
Are you a music professional? *Yes / No*
If you answered Yes to the previous question, please describe your profession.
Do you participate in any professional or amateur music organizations/groups? *Yes / No*
If so, please explain.

Absolute Pitch Abilities – *We like to know how participants classify their own absolute pitch abilities.*
Have you ever attempted to learn absolute pitch? *Yes / No*
If so, using which method?
Do you have Absolute Pitch? *Yes / No/ Don't Know*
(If you do not have absolute pitch, please skip the next 8 questions or click "here" to continue with the survey.)

How do you rank your ability to identify the pitch of a note?
Right about half the time / Mostly Right, but more than a few errors / Rarely Miss
How rapidly can you identify the pitch of a note?
Instantaneously / In 3 seconds / In 15 seconds / In 30 seconds / In 1 minute
At what age did you know that you have Absolute Pitch?
Always / By 5 / By 12 / After Age 15
On which instrument(s) can you determine the pitch without an external reference?
Piano / Violin / Voice / Flute / Any Instrument other instruments?
Can you produce any requested pitch vocally without a reference? *Yes / No*
Have you noticed any changes in your pitch-naming abilities with time? *Yes/No*
If so, when did you first notice the changes?
Please describe the changes:

Family History – *In order to determine how absolute pitch is inherited, we would like to know if any of your family members have absolute pitch abilities.*
How many brothers and sisters do you have, if any? __brothers __sisters
Do any of your family members possess Absolute Pitch? *Yes / No / Don't Know*
If you answered Yes to the previous question, what is their relation to you? (indicate quantities next to all that apply)
Mother / Father / Sister / Brother / Daughter / Son / Aunt / Uncle / Maternal cousin / Paternal Cousin / Nephew / Niece / Grandchild / Maternal grandparent / Paternal grandparent other family?
If you are a twin or multiple birth, do(es) your multiple birth/twin sibling(s) have absolute pitch?

Subjects of Further Study – *Other interesting observations about absolute pitch have been reported in the scientific literature, and we would like to follow up on these in our study population.*

Synesthesia - Some individuals with absolute pitch report having synesthesia (cross-sensory perception, e.g. perceiving notes as color).
Do you have synesthesia? *Yes / No / Don't Know*
If you answered Yes to the previous question, please describe what you perceive:

Neurological Conditions – Some reports have suggested an increased incidence of absolute pitch in people with certain neurological conditions, such as the Autism Spectrum Disorders, Attention Deficit Disorder or ADHD, and Williams Syndrome. We would be interested to learn the prevalence of these conditions in our study population and their family members.
Do you or any family members have these or other neurological conditions? *Yes/No/Prefer not to answer*
If you feel comfortable doing so, please elaborate about which family members possess neurological conditions and which conditions they possess:
Would you be willing to be re-contacted to participate in a follow-up questionnaire if you or any family members possess these or other cognitive abilities or disabilities? *Yes/No*

Medication-Induced Pitch Perception Changes – There have been reports that certain medications (such as carbamazepine or trimipramine) can reversibly shift pitch perception by a semi-tone or more. Absolute pitch possessors are uniquely suited to notice these sorts of changes.
If you have noticed changes in your pitch perception while taking a medication, would you be willing to be re-contacted to take a follow-up questionnaire about your experiences? *Yes/No*

Cycling of Pitch Perception with Hormone Levels – Several small studies have observed that pitch perception by absolute pitch possessors appears to fluctuate with hormonal cycling. This has been observed in both men and women. In women, it appears to change throughout the menstrual cycle and also during pregnancy. These changes are rather small, and are probably not detected by our current pitch naming test.
Would you be willing to be re-contacted to potentially participate in a follow-up study investigating these fluctuations? *Yes/No*

Additional Comments
Please use this space to add any information about yourself or your pitch perception that you think would be useful for our study.

Appendix B. Primer sequences used for candidate gene re-sequencing. Touchdown PCR reactions were performed using Platinum Pfx DNA Polymerase (Invitrogen) often using 2X reaction buffer and 1X or 2X enhancer.

Primer Name	Sequence
ADCY8_5UTR1_F	TCCAGTAGGGTGAGGCTGAT
ADCY8_5UTR1_R	CACCTCCTCCACCAAGACTG
ADCY8_5UTR2_F	AGAATGCTTCTTGGGCTGAA
ADCY8_5UTR2_R	GGGCCTGGAAGTTAAGGGTA
ADCY83primeDelF	TCCAGGGAAATCCTAGCAAA
ADCY83primeDelR	TCAGGTCTTATGGGTCACCTC
ADCY83primeF	TGGAAGAGTGGGACATGTGA
ADCY83primeR	CCTGGCAATCTTGTGGTCAT
ADCY83UTR_2F	TGATTTGGGCACACTCATGT
ADCY83UTR_2R	GCTGCTGAAGGGATTGTTGC
ADCY83UTR_3F	TGTGCCAACATAGCAAGGAA
ADCY83UTR_3R	ACCACAACATGGCAGACAGA
ADCY83UTR4F	AGGGTTTTGCCATGTTGCT
ADCY83UTR4R	GGACTCTGCAAACTGCAATG
ADCY8Ex1_1F	ATTTGGGCTTTCATGGGAC
ADCY8Ex1_1R	CACTGCGATCCAGCGTTC
ADCY8Ex1_2_2F	GGCGAAAAGGAAACCCTGTA
ADCY8Ex1_2_2R	TAGTTCGGGAGCAAGGACTG
ADCY8Ex1_2F	CTACGAGTCCTGGCTCCG
ADCY8Ex1_2R	GGGTGTGGGGACTCTCG
ADCY8Ex1_3F	AGGGGATGAGAACTGTGTGC
ADCY8Ex1_3R	GTCCTTGCGTGCTGCTCTC
ADCY8Ex1_4F	GAGCCTGAGCCCAGGAAG
ADCY8Ex1_4R	AGGTGCAGGAAGCCCAG
ADCY8Ex1_5F	ACCTGCGGCACCAAAGTC
ADCY8Ex1_5R	GAGGCAACCCTGGCTCTC
ADCY8Ex10F	CAGGAATAGAATTTCAGTGTTGC
ADCY8Ex10R	GGCTCCATAAACTTCATGCTC
ADCY8Ex11_2F	TCCTAGGAAATGCCCATCAC
ADCY8Ex11_2R	GAGGGGACAGGAAAGAGGAC
ADCY8Ex11F	TCCATTTCCAGCACTCACAC
ADCY8Ex11R	TTTCTATGGCCCTTCAGCC
ADCY8Ex12_2F	TTAGGATGGCCTGTGTAGGG
ADCY8Ex12_2R	CCACAATGCTGTGGATATGG
ADCY8Ex12_3F	CAGTAAGCCGAGATGATACCAC
ADCY8Ex12_3R	CAGCCCTGTGGATACCAAAA
ADCY8Ex12F	TGGCTAGGAGAGGGAACTGC
ADCY8Ex12R	AGCGGGCATCATTTTCC
ADCY8Ex13F	CCCTTAGAGCAGCCACTATCAC
ADCY8Ex13R	CTGCAAAGAGCAGCAGAGG
ADCY8Ex14F	GTGCCTGAAGCCTACACCTC
ADCY8Ex14R	ACTGCACATCACCCACCAC
ADCY8Ex15F	ATCACTGCAGAACCGACCTC
ADCY8Ex15R	AACGCTTTGACAACGATGC
ADCY8Ex16F	CAGGCTGCTGGAAGAATAAAC
ADCY8Ex16R	GTGACAGAGGCTTCACCACC
ADCY8Ex17F	TGCTGCAGAAATGTTAATGTTC
ADCY8Ex17R	AGGGTCTGTTGGAGCAAGG
ADCY8Ex18F	GTCCAAACACCCCATTGTCC
ADCY8Ex18R	GAAGGGAACATTTGAAGAGAATC
ADCY8Ex2F	GGATTACTACATTAGGATTTACCTTGG
ADCY8Ex2R	AGTCAGGACGTGTTTGGGAG
ADCY8Ex3_2F	GCTCCACCAGAGTCAGAACC

ADCY8Ex3_2R	GTTTGCCTTTCCCACTCAAA
ADCY8Ex3F	CCACCCTCCCATCTGAGAC
ADCY8Ex3R	AGCACCCAAACACACATGG
ADCY8Ex4_2F	CTTCCTGAGCTCCTGAGTGG
ADCY8Ex4_2R	GCTCTGATCAAGACAGCTGGA
ADCY8Ex4F	TATCTGCTGGGATGGTGTTG
ADCY8Ex4R	ACATTGGGGAAGAAGGCTG
ADCY8Ex5F	CTGCTCGCCTCTTTCTCAAG
ADCY8Ex5R	CCTGTGATCGTGCACATTG
ADCY8Ex6F	AAGAAAGTGCAGAGCCAACG
ADCY8Ex6R	CACTGTCACAAAATAAGCCAATG
ADCY8Ex7F	TAGCCTTTGCAGGCAACATC
ADCY8Ex7R	TCTGGATTGGAGATGCACAC
ADCY8Ex8F	CTGAATACAACTGCATGGAATG
ADCY8Ex8R	AAGAAACCAAAACCACAGCC
ADCY8Ex9_2F	GGCTCTCACTGGAAAGTTGG
ADCY8Ex9_2R	CCTGGGTTCAATCTAGGCAGT
ADCY8Ex9F	CCTCAAACACTGCACTCTGC
ADCY8Ex9R	TCAGCTGGCTGACTCCAC
ADCY8intron10F	CCCACATTGTCATGGTTTCA
ADCY8intron10R	GCTGCAGGAGATTTTGTGGT
ADCY8intron11_1F	GGCATTTCCTTCCCTTTCTT
ADCY8intron11_1R	CACATGCATGCACTGATTGA
ADCY8intron11_2F	ATTGATCTCCCAAGCCAGAA
ADCY8intron11_2R	GGGGTAAATCCCTGTACTCCA
ADCY8intron12F	GGTCCACAAATTATTGAACCAG
ADCY8intron12R	CCACCCAGAGAAGACAATGG
ADCY8intron13F	TGAGGGATACACCAGCATGA
ADCY8intron13R	GCTTCAGGCACAGACAACTG
ADCY8intron14F	CAGACAACTGGAGTGGAGCA
ADCY8intron14R	GGGAAACTGGGGGATACAGA
ADCY8intron15_1F	CTTGACACAAGTGGGCCTTC
ADCY8intron15_1R	GGGTCTGGAGGATATGATGC
ADCY8intron15_2F	TCTCACCATCAATGCTGCTT
ADCY8intron15_2R	GACTGGGGTTCAAATCTTGG
ADCY8intron15_3_2F	AAGGGAGAGGTGCACTTATGAT
ADCY8intron15_3_2R	TAGCCATCCCCATCAGTCTC
ADCY8intron15_3F	CTGGAAATGCACTGCGAATA
ADCY8intron15_3R	CAGCTCCAGCCTTCTTCTTG
ADCY8intron15F	AAAGTTACCTGCAAGGGCATT
ADCY8intron15R	TTCATGTCCTCACAGACTCTCA
ADCY8intron17_2F	TTTGTGTTTTGGGTCTGCTG
ADCY8intron17_2R	TGTGGTGGGAAAGGAAAGAA
ADCY8intron17F	TCAATGCTGGTGGTCAATGT
ADCY8intron17R	GCTGCTAAGGGAGAAGCAGA
ADCY8mRNA1F	CCAGGATTTGCGGACTTTTA
ADCY8mRNA1R	CCTCGGTAATCAAAGGCAAA
ADCY8mRNA2F	AGAGCAGCACGCAAGGAC
ADCY8mRNA2R	CGCTGGGTGACAACAAAGT
ASAP1mRNA10F	TCAGCTGTTATTGGAAAGTGATTT
ASAP1mRNA1F	CCATCCAAAAGCAAGCATCT
ASAP1mRNA1R	CGAGGCTACCCTCATACTGG
ASAP1mRNA2F	TCTGATGTGACGGCTGAGAC
ASAP1mRNA2R	CAGATGAGAATTCGGCATTTAAC
ASAP1mRNA3F	AGAGAGCACGCAAAACAACA
ASAP1mRNA3R	GGCTTTTGTCAGGTCTTCCA
ASAP1mRNA4F	CCTTCTCACCTGCCAAGTAAA

ASAP1mRNA4R	CCGTCTGCTTATCCAGGTTC
ASAP1mRNA5F	AAGGGGTAGAGCTAATGGAACC
ASAP1mRNA5R	CCTTAGTGCCACTTTCTGAGG
ASAP1mRNA6F	AGTTCCTTGGGGTAACGATG
ASAP1mRNA6R	ACACTGGAAAGACCCCCTTC
ASAP1mRNA7F	GTGAGGCGAGTGAAGACCAT
ASAP1mRNA7R	TTTGGGTCAGTGAATATTTTGC
ASAP1mRNA8F	GGGTTCTGTTTTCCTGGTGA
ASAP1mRNA8R	CAAAGAGCATGGGAAAAGAAA
ASAP1mRNA9F	AGCATCCATTGCATCCATTT
ASAP1mRNA9R	CAGACTTCCATCCGGACACT
ASAP1RNABT1F	GTGACGGCTGAGACATGAGA
ASAP1RNABT1R	CAGCAGATCCACACCCTTTT
DDEF1Ex1F	TCCCAGACTTCCTGTGCTCT
DDEF1Ex1R	GCTGGTGTCACTGTTGTTGC
DDEF1Ex23F	CTCAGAGAGCCACAGGAACC
DDEF1Ex23R	CCTTGGGCTGGTCAGTGTAT
DDEF1Ex29F	GTGTCAGGGCACATGTGAAT
DDEF1Ex29R	CAGATGAGAATTCGGCATTTAAC
DDEF1Ex4F	TCCTAACCTAACCTTTGGGACA
DDEF1Ex4R	GCAGTTTCTCTCTCAGCTTCG
EFR3AEX10F	TGAGTTGCTCATAAATGCCC
EFR3AEX10R	CTCATGCTTGCTTGTTTTCG
EFR3AEX11F	TTGTGAGGGATTGTTTTACCAC
EFR3AEX11R	AAGCTGAAAATGATTCTAAAGGC
EFR3AEX12F	TTGGCATGTGAAGTGAGCC
EFR3AEX12R	AAAGAACGGTTGTAACATGAAGC
EFR3AEX13F	TTTTCTCCAGGCTTCTGTCC
EFR3AEX13R	AAGAGTACATACGTGCCCTCC
EFR3AEX14F	GGGCACGTATGTACTCTTCAG
EFR3AEX14R	GAAGATGGCGTCTTGAGGTC
EFR3AEX15F	AAAAGTATATGCATGTGAAATCTGC
EFR3AEX15R	TCATGTCCTTGGATAACGGC
EFR3AEX16F	CCAAACTACCTGCAGTTTTCC
EFR3AEX16R	GCTGTAGTTTCAGTTTACTGGCTG
EFR3AEX17F	AACAAACTTTACATTTGGGCTG
EFR3AEX17R	TGATGATTAGTATAAAGCACCAGG
EFR3AEX18F	GAGGCTTTTGCTCTGGAATG
EFR3AEX18R	AATTTTCATTGTTAACTGGGTAATG
EFR3AEX19F	AAAAGTCCGAGAACACGCTG
EFR3AEX19R	GACAGGCTTTGTATAAAGTTTCTTAGC
EFR3AEX1F	GCTCGACAAGGGATCCTG
EFR3AEX1R	AGGAGTACCCGGCCCAAG
EFR3AEX20F	TCACAGCTCAGACCCAACAC
EFR3AEX20R	CCCTCTACGTATTCTGTCATCG
EFR3AEX21F	AAATTGCCCTTAATTTTGGATG
EFR3AEX21R	GCTTGGAACTTATCAACTAATCACC
EFR3AEX22F	CATTGAGACGAGGAAAGGG
EFR3AEX22R	TGAACCACAAATTCATAGGGG
EFR3AEX23_1F	GGAGTCTGACTTTGATATTCGC
EFR3AEX23_1R	GCCTTTGCAGAAAACAACAAC
EFR3AEX23_2F	CATGTTTTGGTTTCACTTTATTCC
EFR3AEX23_2R	TCTTCACAGGTGAACTAATAGCTG
EFR3AEX23_3F	TTAGTTTACAGGCTGTGCTTTG
EFR3AEX23_3R	AAACATCCCATTCTTCACTGC
EFR3AEX23_4F	GCTACCAAGATCACAGGTGC
EFR3AEX23_4R	GAGCAAAGTTTGTGTTTCACG

EFR3AEX23_5F	CTTCACCAGTCCGTAAAGCC
EFR3AEX23_5R	CCTGAACTGGCTACAACACC
EFR3AEX23_6F	TGGATCTTTTCACTGGCTGAC
EFR3AEX23_6R	GGGCTCAAAAGAAGCAATTC
EFR3AEx2F	TCAGAAACTTGAATTTCCAGGC
EFR3AEX2R	AACCTTAAGGAGGTCAGCGG
EFR3AEX3F	AAGGGAAATTGTAAAGCAGGC
EFR3AEX3R	TTGTCTCACTAAATGGCAAAATAATC
EFR3AEX4F	TTTCAGTTCTTCCAATGTAAAGC
EFR3AEX4R	TGCCCTATTCAGGGAATTTG
EFR3AEx5F	CAAGACCCTATCTCAAGAAAGAAAG
EFR3AEX5R	ATGGAAAGCCGTATTTCAGG
EFR3AEX6F	TTCTCTTAGCTACGCAAACAGTC
EFR3AEX6R	GAATAATAATCTGTAGTCTGCTAAGCC
EFR3AEX7F	GAGTGTATGATTCATTTTGGTGG
EFR3AEX7R	TGAACCACTGGTCAAACACATC
EFR3AEX8F	ATTAAGCTTTTCCTATTGTTGAGG
EFR3AEX8R	ACAGCTTCCCACATCTGTTTG
EFR3AEX9F	TGGTCATTTACGTGTCAGGTG
EFR3AEX9R	CACTGCTATGCATACTGAAAGG
ELAVL2AltEx1&2F	GTCCTGGGCTTATACGCAAT
ELAVL2AltEx1&2R	GGGTATCGCAGATGATACCAA
ELAVL2AltEx1_3F	TGCACCTAGGGACAGTGTTG
ELAVL2AltEx1_3R	GTGAGTAGGGCAGCACACAA
ELAVL2Ex1longF	TACCTCCCCGCACAACTTAC
ELAVL2Ex1longR	TCTGCCATGGACATTTACGA
ELAVL2Ex1S&TF	CTGCTCTGACTCCCCGTTAG
ELAVL2Ex1S&TR	GCCGGTGTTAAGTCTCCAAA
ELAVL2Ex1SF	CGAGGCTAAGGTCTGTGCG
ELAVL2Ex1SR	GGTTGCCAGTGCAACACAG
ELAVL2Ex1TF	AGCTGAGCTGCTGAAGCC
ELAVL2Ex1TR	CCACCCCTACCCTGCTAGAC
ELAVL2Ex1UF	CGTCTCTTTTCTTTGTTCCTTGG
ELAVL2Ex1UR	ACAGGTTTTCTGGGTTTGCC
ELAVL2Ex2F	AAGTGTTTAAATTCTTGGATGGAG
ELAVL2Ex2R	TTGCAGCAGTGAATTATTTACAAG
ELAVL2Ex3F	CCATATGCCAAAATGAAAGTTG
ELAVL2Ex3R	TATATGGGCCCAAAAGGAAG
ELAVL2Ex4_2F	CTGTTCATGGGAAAGTTACCG
ELAVL2Ex4_2R	AAATGACTCAAGCACGCTCA
ELAVL2Ex4F	TTCTAAAGGGAGGTGGGTG
ELAVL2Ex4R	ACAGAATATTTCACAATCTGCTAGTC
ELAVL2Ex5F	AAAATTAATGTGTCCTTTCTTCTCTC
ELAVL2Ex5R	AAAGCAAACCAGAGATCCTGTC
ELAVL2Ex6U_2F	CAGAGCTTTCGCACTTCCTC
ELAVL2Ex6U_2R	GGCATCCGGTGGTATAAAAA
ELAVL2Ex6UF	GAGCCCAACTTTCTTGAACATC
ELAVL2Ex6UR	CCCAAAATCAAAGAAACCAATC
ELAVL2Ex7_1F	TAGCATGTCACTGCAAAGCC
ELAVL2Ex7_1R	CCTTCCCCAACCCAAATC
ELAVL2Ex7_2F	AACTGCCTTGAACCTGTGAG
ELAVL2Ex7_2R	TTCAAACAATACTGCAAGTACATCAC
ELAVL2Ex7_3F	TTTAAATTACCGAGAGATGGGG
ELAVL2Ex7_3R	TGCCTGTTCTAAGGGGAAG
ELAVL2Ex7_4_2F	AAAGCTGAAGGCAAAAATGC
ELAVL2Ex7_4_2R	CCCAGCCATAAAATGGTGTC
ELAVL2Ex7_4F	CCATAGGTTTTGAACAAATTTCC

ELAVL2Ex7_4R	TCAGAAGGTGTCATTCAGTCC
FAM49BEx12F	CGCTGTCTTGCTCATAGCTG
FAM49BEx12R	TTTCCACCTTCCAGATAAAACAT
FAM49BEx15F	TGTACAAATGGTACCCAACACAA
FAM49BEx15R	TCTCCTGGGAAGTCCCTCTT
FAM49BEx1F	GACCACCTGTCTCTGGCTTC
FAM49BEx1R	AGGGGATGAGGAGAAAAAGC
FAM49BEx8F	GCCAAGGTGGGTGGATCA
FAM49BEx8R	TCTCCATCGCTACAGTAAGTCC
FAM49BmRNA1F	TTGAGTGATGCCACAACAAAA
FAM49BmRNA1R	CAGCAGGTGCTTATTCCAGA
FAM49BmRNA2F	CAGGAGCGGGACTGAGAG
FAM49BmRNA2R	ACGCACACATCACAGCTTTT
FAM49BmRNA3F	GAGGTCTTCTGGGAGCCTTA
FAM49BmRNA3R	TGGCAGGATTTGTCATCTTG
GDAP1-1-F	CAAAAGGTGCGCTTGGTAAT
GDAP1-1-R	GGGATCATGGAGTCCACAGT
GDAP1-2-F	CAGTGTGGGAGGGAGAAGTC
GDAP1-2-R	GGCCGCTATTCATTTTAACC
GDAP1-3a-F	CAGTGTGGGAGGGAGAAGTC
GDAP1-3a-R	CGTAATAGGTTTCCAAGTTTGGTC
GDAP1-3b-F	GAAGAAATGAAGAAACCCCAGA
GDAP1-3b-R	AGCAATGTGTGTGATTCATAAGC
GDAP1-3c-F	GCTGCCTGTCTCATTGGTAA
GDAP1-3c-R	TCGGCACATCATCTCTATGC
GDAP1-3d-F	CTTGGAACTGCAACAAATGG
GDAP1-3d-R	TGAGGTCATTACTATCTTTGCTTGA
GDAP1-3e-F	TTTCCTTGTACAGTTTCTTTGGAA
GDAP1-3e-R	GGCCGCTATTCATTTTAACC
GSDMCmRNA1F	TGGCTCAGCTCTCAAAGGAT
GSDMCmRNA1R	TCCATCCTAAGGCCACACTC
GSDMCmRNA2F	AGGCAGTTCAGAGGTGCTTC
GSDMCmRNA2R	TGAAACGCTTACGTCTTTAACA
GSDMCmRNA3F	CTTCAGATACCCCTGGAGCA
GSDMCmRNA3R	AAGATCAGCCAGGCCAAGA
GSDMCmRNA4F	GCCACCATTCCTGGCTAA
GSDMCmRNA4R	CTGAAGAGTCAGCGCCTTCT
GSDMCmRNA5F	GTGGTGAVAGAGGCTGTTGA
GSDMCmRNA5R	TCCTTTGGGTTAAACCATGC
KCNB2-1-F	CGTTTGGCCAAGAACTTGAT
KCNB2-1-R	GCTAATTGGCGGTTGTCATT
KCNB2Ex1F	CTCCGCCACAGACACACA
KCNB2Ex1longF	CTGCGTCCCTCTAGTCCAGT
KCNB2Ex1R	TGGCAGTTTCCAATTCCTTT
KCNB2Ex2_2F	CTGGAAGTGTGCGACGACTA
KCNB2Ex2longF	TTGTCTTTTCCCCCTTTCCT
KCNB2Ex2R	CCAGGCCCCTTTCTACAGAC
KCNB2Ex3_1F	ATCACCGACCCATCTGTCTC
KCNB2Ex3_1R	CCCATGGCCAGAAACAATA
KCNB2Ex3_2F	TCTGGGTTTCACCCTTAGGC
KCNB2Ex3_2R	GGAGCTTGTTTCCGACAGAG
KCNB2Ex3_3F	GAGAGTCCGCCAACACAAAG
KCNB2Ex3_3R	CCCTAGCTCTTTGGTGCTCTT
KCNB2Ex3_4F	TCTCACTTGCACATGAAGTTCC
KCNB2Ex3_4R	CAGTTTCTGCTTGGGAAACC
KCNB2Ex3_5F	AGGGAGACAGACCCTTGCTG
KCNB2Ex3_5R	CAGGCATTATTAACCGCATTT

KCNB2Ex3_6F	CAACCCAGGAGACACAGGTT
KCNB2Ex3_6R	TCCAGGCCCCATTATTTACA
KCNB2intron1F	TGTTGGAGGCCTTAAGCAAT
KCNB2intron1R	CAGGCATTTCCACTGACAAA
KCNB2intron2F	AATCCCCACTTGGGCTATCT
KCNB2intron2R	AAAGCAATGGTGGAAAGCAC
LAMB1RNA1F	GGCAATCCATCAGAAGTTGG
LAMB1RNA1R	GCGCCAGGATGTCTTTTATC
LAMB4RNA1F	ACTGCCAGCACAACACTGAG
LAMB4RNA1R	AGCTGCAGACTGGGTTTCAT
NRCAMmRNA10F	ACCACTCTGGACAGCGTCTC
NRCAMmRNA10R	TTCTCCAGAATGTCCCATGA
NRCAMmRNA11F	ACATCTGTGGTTGTGGCAAA
NRCAMmRNA11R	GGGCCTAATTCATGTGTGCT
NRCAMmRNA12F	AGTGCTCCCTCGTCTTTGAA
NRCAMmRNA12R	GGCCTGTCTCAAACACATCC
NRCAMmRNA1F	TGCCAGGAACAGCATACAAA
NRCAMmRNA1R	GCCGGCTCTTTCTCTTTCTT
NRCAMmRNA2F	GAGCAGCCAGAGGGGATAG
NRCAMmRNA2R	TTGAGGGAAATCCAGTCACA
NRCAMmRNA3F	CCCTGAAATCCAGCCTATGA
NRCAMmRNA3R	CGGAGTAACAGGAGCCAAGA
NRCAMmRNA4F	TGGACATTGTTGTGAAATTGG
NRCAMmRNA4R	AAGCATACAACAGCATGCCA
NRCAMmRNA5F	CAATGCTGGGAAAAGAAGGA
NRCAMmRNA5R	TCAAACATGCCAGGTTACTAGG
NRCAMmRNA6F	TGAATGGGATTGGAAAGCAT
NRCAMmRNA6R	TTCCCTTCGCTCATGATGTT
NRCAMmRNA7F	ATCCAGTGTGAAGCCAAAGG
NRCAMmRNA7R	ACTTGCATTGCCTTCTGGAG
NRCAMmRNA8F	TCAAACCATACAGCAGAAGCA
NRCAMmRNA8R	GACCCAAAGAAGGCACAGTC
NRCAMmRNA9F	TCAGTGCAATGCCTCTAATGA
NRCAMmRNA9R	TGGGGCTATTGTTGTCATCG
NRCAMRNABT1F	GAGCAGCCAGAGGGGATAG
NRCAMRNABT1R	TTCCCTTCGCTCATGATGTT
NRCAMRNABT2F	GGTCTAATGCCAGGAACAGC
NRCAMRNABT2R	TAGGGGTTCAGCTGAGGGTA
NRCAMsplicingF	AGGAAGCAGTAACAACTGTGGA
NRCAMsplicingR	CAAAGAAGCTCCGAGAACCA
PIK3CGmRNA1F	CCAAAATTCAGCAAAGCACA
PIK3CGmRNA1R	GGCATCCCGGATATATTCAA
PIK3CGmRNA2F	GGCATGGAGCTGGAGAACTA
PIK3CGmRNA2R	CCAGGCTGGAGTGTAAGGAC
PIK3CGmRNA3F	TGCCTTATCCATTTCCCATT
PIK3CGmRNA3R	GCTTTCGGGAATATCCATCA
PIK3CGmRNA4F	TTGCCAACAACTGCATCTTC
PIK3CGmRNA4R	ATCGGGTGGCAGTAATTGTC
PIK3CGmRNA5F	GCCGTGGAGAATACGTCCT
PIK3CGmRNA5R	TCTGGACTGGGCTATCTCAC
PIK3CGmRNA6F	GCTTGGAGGACGATGATGTT
PIK3CGmRNA6R	TGTGCTTTGCTGAATTTTGG
PIK3CGmRNA7F	GCGCCAAGACATGCTTATTT
PIK3CGmRNA7R	TGCACAGTCCATCCTTTGTC
PIK3CGmRNA8F	GCCCCAGTTAACAAGCAAAG
PIK3CGmRNA8R	AGACCAGCAGAGGAAGGTCA
PIK3CGmRNA9F	TGGTGTGCTAAAAGCAAGGA

PIK3CGmRNA9R	TTTTCAGGAAATCTGAAGGATG
PKIA-1-F	CAAGTGGCAACAGCAATGAA
PKIA-1-R	CCAAACACAAGCCACATGAT
PKIA-2-F	TGGTAGCAATGACTGATGTGG
PKIA-2-R	TGAACCATTGGCATATTACTGG
PKIA-3a-F	TGGTAGCAATGACTGATGTGG
PKIA-3a-R	GGGTGAACACTTGAGCCTGA
PKIA-3b-F	GCTCCGACCTAGATGATGATTC
PKIA-3b-R	TTCCAATTGTCCAAAATTCTCA
PKIA-3c-F	TTCATCAAGACTCCATTGCTTT
PKIA-3c-R	TGAACCATTGGCATATTACTGG
STAU2-1-F	AAGCACTGCAGAATGAACCT
STAU2-1-R	CTAATAGGGTTCATCCCTTGG
STAU2-2-F	GAGCCGTCTGCAAAGTGTC
STAU2-2-R	GAAAGCCTTGAATCCTTGCT
STAU2-3a-F	GAGCCGTCTGCAAAGTCTC
STAU2-3a-R	TCATTCTGCAGTGCTTGGAG
STAU2-3b-F	GGGGCATGTACAATCAGAGG
STAU2-3b-R	CGAGGCATTCCTCTTTCTGA
STAU2-3c-F	AGTAAAGGCCGGACCAGAAT
STAU2-3c-R	GAAAGCCTTGAATCCTTGCT
TCEB1-1-F	CCTGGGGAAGCAAAGTAGAA
TCEB1-1-R	CGGTGGAGCTGTTAGTGTAGC
TCEB1-2-F	AACTACTAAAGTTCCTGGGGAAGC
TCEB1-2-R	TTTCTCAGTTTGTGAAAATGTCC
TERF1-1-F	GCTTGCCAGTTGAGAACGAT
TERF1-1-R	CCATCATGTGGTTGTAGCTGA
TERF1-2-F	ATTTAACATGGCGGAGGATG
TERF1-2-R	GCTGAAATTGCGCCACTG
TERF-3a-F	ATTTAACATGGCGGAGGATG
TERF-3a-R	AAAAAGGAATGAAATGTATCTTTCTGA
TERF-3b-F	TGTTTGTATGGAAAATGGCAAC
TERF-3b-R	GACACTTGTCCGGTTGTTGA
TERF-3c-F	CAGCCGGTAACTCCTGAAAA
TERF-3c-R	GCTGAAATTGCGCCACTG
TUSC1_e1_F	CTCCTCCGTTCCCAGCTAC
TUSC1_e1_R	TAATACTCGCCGCAACCTTT
TUSC1_e2_F	TGCCTATATCGTGATCTTTTGAA
TUSC1_e2_R	ACAGCCTTGTTGCTCACAAA
TUSC1_U1_F	CTTGATCCCAGAAACTTTGGA
TUSC1_U1_R	GACTGAGGCGTGTCCTGTCT
TUSC1_U2_F	TCAGCTGAGGATACACGCTTT
TUSC1_U2_R	CAAAGCCTGACGGAAGAGG
TUSC1end1_2F	GCCTCTACAGGAACCCGACT
TUSC1end1_2R	AAAATTCTCTGGGGGCAAAC
TUSC1end2_2F	TCGGTGTTACTTGGATGTCCT
TUSC1end2_2R	CCCTTCACTGACTCTTTTCG
TUSC1Ex1_1F	CTCCTCCGTCGTCACCC
TUSC1Ex1_1R	GCTGGCCTCTTCAGGGAC
TUSC1Ex1_2F	GGCGAGCCACTTGGAGG
TUSC1Ex1_2R	CCGTAGTCCAAGGTATCCGC
TUSC1UTR1_2F	TTCTTGCTAGGGCATACTTGC
TUSC1UTR1_2R	CTGGCTTAGAGCCAAAGGAG

Appendix C. SNPs found to be present when Sanger sequencing approximately 8 AP probands. Those highlighted in red were not in dbSNP when we discovered them. HapMap CEU allele frequencies are shown for comparison when available. Physical positions are on hg18.

Chr	Bp	SNP	Gene	Location	A1	A2	11	12	22	AP-A1	CEU-A1
7	106300247	rs1129293	*PIK3CG*	Ser675	C	T	4	2	0	83.3%	
7	106333893	rs12667819	*PIK3CG*	3'UTR	G	A	0	4	2	33.3%	54.9%
7	107612125	rs401433	*NRCAM*	Ala735	C	G	6	1	0	92.9%	
7	107621849	rs6958498	*NRCAM*	Pro545Ala	C	G	0	3	4	21.4%	
7	107621970	rs404287	*NRCAM*	Ala534	C	T	0	4	3	28.6%	19.5%
7	107625700	rs381318	*NRCAM*	Val429	G	T	0	1	4	10.0%	25.2%
8	73641739	rs2451035	*KCNB2*	intron	C	T	0	5	3	31.3%	31.4%
8	73641763	rs2383866	*KCNB2*	intron	A	G	0	1	7	6.3%	
8	73642277	rs2451034	*KCNB2*	intron	A	G	1	4	3	37.5%	34.4%
8	74009940	rs11784364	*KCNB2*	intron	C	T	3	2	2	57.1%	59.8%
8	74009948	rs12548517	*KCNB2*	intron	C	T	5	2	0	85.7%	
8	74010009	rs7828692	*KCNB2*	intron	G	A	0	4	3	28.6%	49.1%
8	74010079	rs12545401	*KCNB2*	intron	C	A	5	2	0	85.7%	93.8%
8	74010290	rs35339929	*KCNB2*	intron	G	T	0	2	5	14.3%	
8	74010348	rs67574145	*KCNB2*	intron	A	G	0	2	5	14.3%	
8	74690550	rs949493	*STAU2*	Met166Val	C	T	0	1	6	7.1%	8.5%
8	75439514	rs1135715	*GDAP1*	3'UTR	A	G	4	3	0	78.6%	79.6%
8	75439987	rs4737414	*GDAP1*	3'UTR	C	G	4	3	0	78.6%	
8	75441012	rs6472842	*GDAP1*	3'UTR	A	G	3	4	0	71.4%	
8	75441035	rs16938896	*GDAP1*	3'UTR	G	T	4	3	0	78.6%	
8	75441371	rs10504580	*GDAP1*	3'UTR	A	G	3	4	0	71.4%	
8	130922629	rs873065	*FAM49B*	3' of gene	C	T	4	2	2	62.5%	
8	131021293	rs11785995	*FAM49B*	5' of gene	G	A	6	2	0	87.5%	87.8%
8	131134335	rs4236749	*ASAP1*	3'UTR	C	T	6	2	0	87.5%	82.7%
8	131136030	rs11781272	*ASAP1*	3'UTR	C	G	0	2	4	16.7%	
8	131136110	rs11781294	*ASAP1*	3'UTR	G	T	4	2	0	83.3%	
8	131193741	rs966185	*ASAP1*	Ile728Val	T	C	1	6	1	50.0%	48.2%
8	131295834	rs1469288	*ASAP1*	intron	C	T	4	3	1	68.8%	
8	131295934	rs2303444	*ASAP1*	intron	A	G	5	3	0	81.3%	85.0%
8	131296201	rs1469286	*ASAP1*	intron	G	A	6	2	0	87.5%	80.1%
8	131483138	rs10090106	*ASAP1*	intron	C	T	0	2	2	25.0%	
8	131483265	rs10090231	*ASAP1*	intron	A	G	3	5	0	68.8%	70.4%
8	131483439	rs10090767	*ASAP1*	5' of gene	G	A	3	5	0	68.8%	70.4%
8	131483746	rs11992885	*ASAP1*	5' of gene	A	G	0	5	3	31.3%	
8	131483766	rs11992932	*ASAP1*	5' of gene	G	T	3	5	0	68.8%	70.4%
8	131483814	rs11992957	*ASAP1*	5' of gene	C	G	3	5	0	68.8%	
8	131848488	rs62519397	*ADCY8*	3' of gene	C	T	6	1	0	92.9%	
8	131848680	rs57425225	*ADCY8*	3' of gene	A	G	2	4	1	57.1%	
8	131848842	rs6470848	*ADCY8*	3' of gene	C	A	1	4	2	42.9%	78.8%
8	131849139	rs2572862	*ADCY8*	3' of gene	C	A	5	2	0	85.7%	89.8%
8	131849331	rs6990380	*ADCY8*	3' of gene	G	A	0	4	3	28.6%	45.1%
8	131849409	rs6990427	*ADCY8*	3' of gene	G	C	0	4	3	28.6%	45.6%
8	131849424	rs11997892	*ADCY8*	3' of gene	C	T	0	4	3	28.6%	
8	131849440	rs10097218	*ADCY8*	3' of gene	T	C	0	3	4	21.4%	40.3%
8	131852386	rs17225390	*ADCY8*	3' of gene	C	T	0	2	6	12.5%	
8	131861181-131861241	rs55861470	*ADCY8*	3' of gene	+	-	0	3	4	21.4%	
8	131861401	rs263247	*ADCY8*	3' of gene	C	T	0	2	5	14.3%	36.3%
8	131864442	rs263249	*ADCY8*	intron	A	G	0	2	6	12.5%	36.3%
8	131864555	rs56263895	*ADCY8*	intron	T	-	6	2	0	87.5%	
8	131864593	rs873667	*ADCY8*	intron	A	G	5	3	0	81.3%	
8	131864671	rs873666	*ADCY8*	intron	A	G	5	2	0	85.7%	
8	131874519	rs10956552	*ADCY8*	intron	A	G	5	3	0	81.3%	

Chr	Bp	SNP	Gene	Location	A1	A2	11	12	22	AP-A1	CEU-A1
8	131874785	rs11776881	*ADCY8*	intron	G	T	3	3	2	56.3%	
8	131879758	rs13258256	*ADCY8*	intron	A	G	1	3	3	35.7%	
8	131879803	N/A	*ADCY8*	intron	A	C	0	1	6	7.1%	
8	131880021	rs7015079	*ADCY8*	intron	T	G	0	2	5	14.3%	29.2%
8	131880514	rs34890820	*ADCY8*	intron	C	G	0	1	6	7.1%	
8	131880623	rs57000686	*ADCY8*	intron	C	T	2	3	2	50.0%	
8	131880835	rs16904360	*ADCY8*	intron	T	G	5	2	0	85.7%	90.3%
8	131880862	rs16904361	*ADCY8*	intron	A	T	0	2	5	14.3%	
8	131880924	N/A	*ADCY8*	intron	A	G	5	2	0	85.7%	
8	131880957	rs12547373	*ADCY8*	intron	C	T	0	3	4	21.4%	
8	131880998	rs12545113	*ADCY8*	intron	T	G	5	2	0	85.7%	90.3%
8	131881003	rs7000229	*ADCY8*	intron	A	G	2	3	2	50.0%	
8	131886873	rs17226545	*ADCY8*	intron	T	C	1	2	5	25.0%	35.0%
8	131886953	rs384271	*ADCY8*	intron	C	T	1	1	6	18.8%	
8	131886956	N/A	*ADCY8*	intron	C	T	0	1	7	6.3%	
8	131886957	rs402620	*ADCY8*	intron	A	G	4	4	0	75.0%	
8	131887071	rs1543020	*ADCY8*	intron	A	C	6	2	0	87.5%	86.3%
8	131887220	rs17226587	*ADCY8*	intron	A	G	0	3	5	18.8%	
8	131887277	rs263236	*ADCY8*	intron	C	T	0	1	7	6.3%	
8	131887416	rs263235	*ADCY8*	intron	A	G	1	1	6	18.8%	
8	131895803	rs1435446	*ADCY8*	intron	G	A	5	3	0	81.3%	75.2%
8	131895882	rs263265	*ADCY8*	intron	A	G	0	1	7	6.3%	
8	131896138	rs263266	*ADCY8*	intron	A	G	7	1	0	93.8%	
8	131896244	rs3793389	*ADCY8*	intron	C	T	5	3	0	81.3%	
8	131896303	rs3829031	*ADCY8*	intron	C	T	5	3	0	81.3%	
8	131896313	rs377711	*ADCY8*	intron	C	G	0	1	7	6.3%	
8	131902937	rs55879109	*ADCY8*	intron	A	T	6	2	0	87.5%	
8	131905302	rs6996688	*ADCY8*	intron	C	T	4	4	0	75.0%	77.9%
8	131905535	rs6997439	*ADCY8*	intron	G	T	4	4	0	75.0%	77.9%
8	131905801	rs263263	*ADCY8*	intron	A	T	0	3	5	18.8%	
8	131905912	rs263264	*ADCY8*	intron	T	C	6	2	0	87.5%	73.9%
8	131906087	rs12375420	*ADCY8*	intron	A	G	1	4	3	37.5%	
8	131917647	rs2259296	*ADCY8*	intron	A	G	0	3	5	18.8%	18.1%
8	131922543	rs6993838	*ADCY8*	intron	C	G	7	1	0	93.8%	
8	131922595	rs11780751	*ADCY8*	intron	A	T	4	4	0	75.0%	
8	131922625	rs62518066	*ADCY8*	intron	A	G	0	4	4	25.0%	
8	131922640	rs263256	*ADCY8*	intron	A	T	0	3	5	18.8%	
8	131924894	rs3914070	*ADCY8*	intron	C	T	0	4	4	25.0%	
8	131924906	rs263258	*ADCY8*	intron	A	G	6	2	0	87.5%	74.1%
8	131925161	rs6991158	*ADCY8*	intron	A	G	4	4	0	75.0%	
8	131925225	rs17227830	*ADCY8*	intron	G	A	4	3	1	68.8%	86.7%
8	131925252	rs263260	*ADCY8*	intron	C	T	0	2	6	12.5%	
8	131925303	rs16904374	*ADCY8*	intron	G	A	4	4	0	75.0%	79.6%
8	131929678	rs3793393	*ADCY8*	intron	C	T	0	5	3	31.3%	
8	131929803	rs12543363	*ADCY8*	intron	C	T	0	4	4	25.0%	
8	131929989	rs12548296	*ADCY8*	intron	C	T	3	5	0	68.8%	70.5%
8	131929994	rs7820412	*ADCY8*	intron	A	C	3	5	0	68.8%	
8	131930066	rs174493	*ADCY8*	intron	C	T	6	2	0	87.5%	73.2%
8	131930184	rs12548835	*ADCY8*	intron	C	T	3	5	0	68.8%	70.5%
8	131948841	rs12544368	*ADCY8*	intron	C	T	1	4	3	37.5%	58.0%
8	131948991	rs34633756	*ADCY8*	intron	T	-	7	1	0	93.8%	
8	131949070	rs9694427	*ADCY8*	intron	C	T	0	1	7	6.3%	
8	131949474	rs7838092	*ADCY8*	intron	A	T	3	4	1	62.5%	
8	131949494	rs7838097	*ADCY8*	intron	A	G	7	1	0	93.8%	
8	131966264	rs4128982	*ADCY8*	intron	A	G	1	0	7	12.5%	31.0%
8	131991138	rs12547243	*ADCY8*	Pro546	A	G	1	3	4	31.3%	27.4%

Chr	Bp	SNP	Gene	Location	A1	A2	11	12	22	AP-A1	CEU-A1
8	131991209	rs12545028	*ADCY8*	Arg523	T	G	5	3	0	81.3%	84.5%
8	132033482	rs56152625	*ADCY8*	intron	A	G	0	1	7	6.3%	
8	132071758	rs1329803	*ADCY8*	intron	G	A	3	3	2	56.3%	51.8%
8	132071952	rs13278912	*ADCY8*	Leu327	G	A	7	1	0	93.8%	87.5%
8	132071996	rs11991124	*ADCY8*	intron	A	G	0	1	7	6.3%	
8	132121524	rs2228949	*ADCY8*	Ala80Thr	C	T	7	1	0	93.8%	
8	132122594	rs913818	*ADCY8*	5'UTR	A	G	0	1	7	6.3%	
8	132123334	rs3829210	*ADCY8*	5'UTR	G	A	3	3	2	56.3%	49.6%
8	132124848	rs55681582	*ADCY8*	5' of gene	C	T	0	2	6	12.5%	
8	133049891	rs6471017	*EFR3A*	intron	G	A	0	4	4	25.0%	15.0%
8	133052006	rs1051221	*EFR3A*	Asn365Asp	A	G	4	2	2	62.5%	48.2%
8	133057660	rs2270876	*EFR3A*	intron	A	T	2	2	4	37.5%	
8	133065781	rs16904564	*EFR3A*	intron	A	C	4	2	2	62.5%	66.8%
8	133066535	rs10085968	*EFR3A*	intron	C	T	1	2	5	25.0%	
8	133078005	rs2270873	*EFR3A*	intron	G	A	0	3	4	21.4%	15.0%
8	133085135	rs16904572	*EFR3A*	intron	A	G	7	1	0	93.8%	
8	133092345	rs4736529	*EFR3A*	3'UTR	G	C	0	4	4	25.0%	15.0%
8	133092662	rs72631809	*EFR3A*	3'UTR	C	T	0	1	7	6.3%	
8	133092777	rs3783568	*EFR3A*	3'UTR	A	G	0	4	4	25.0%	
8	133094020	rs1051257	*EFR3A*	3'UTR	A	T	0	3	3	25.0%	
9	23680034	rs72631808	*ELAVL2*	3' of gene	A	G	7	1	0	93.8%	
9	23681051	rs9696499	*ELAVL2*	3'UTR	C	G	0	2	6	12.5%	9.6%
9	23681206	rs72631807	*ELAVL2*	3'UTR	C	T	0	1	7	6.3%	
9	23682125	rs41271151	*ELAVL2*	3'UTR	C	G	5	3	0	81.3%	
9	25666579	N/A	*TUSC1*	3'UTR	A	G	0	1	6	7.1%	
9	25666584	rs4592123	*TUSC1*	3'UTR	G	T	5	2	0	85.7%	
9	25666809	rs10812298	*TUSC1*	3'UTR	G	T	1	3	3	35.7%	
9	25666895	rs10812299	*TUSC1*	3'UTR	A	G	3	3	1	64.3%	
9	25666913	rs7044566	*TUSC1*	3'UTR	A	T	2	2	3	42.9%	
9	25666955	rs7028310	*TUSC1*	3'UTR	C	G	0	3	4	21.4%	
9	25667215	rs12348	*TUSC1*	3'UTR	T	C	3	2	1	66.7%	58.0%
9	25667257	rs1128957	*TUSC1*	3'UTR	C	G	2	2	2	50.0%	48.2%
9	25667349	rs1128953	*TUSC1*	3'UTR	A	C	2	3	3	43.8%	45.6%
9	25667588	rs10812300	*TUSC1*	3'UTR	G	T	3	3	2	56.3%	
9	25667698	rs72631815	*TUSC1*	Ser207Ala	G	T	1	5	2	43.8%	
9	25667933	rs35110225	*TUSC1*	Ala129	A	G	2	1	4	35.7%	
9	25667953	rs34498078	*TUSC1*	Asn123Asp	A	G	0	1	7	6.3%	
9	25668122	rs72631814	*TUSC1*	Ala66	C	G	2	3	1	58.3%	
9	25668196	rs72631813	*TUSC1*	Ser41Gly	A	G	6	1	0	92.9%	
9	25668639	rs61483294	*TUSC1*	5'UTR	C	G	6	1	0	92.9%	
9	25668640	rs10967034	*TUSC1*	5'UTR	A	T	4	2	1	71.4%	
9	25668797	rs34772164	*TUSC1*	5'UTR	C	T	5	2	0	85.7%	
9	25668887	rs60018547	*TUSC1*	5' of gene	C	T	1	5	1	50.0%	
9	25669024	rs10738727	*TUSC1*	5' of gene	G	C	1	4	2	42.9%	43.8%
9	25669137	rs10738728	*TUSC1*	5' of gene	A	G	1	1	1	50.0%	
9	25669141	rs10738729	*TUSC1*	5' of gene	A	G	1	1	1	50.0%	

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Elizabeth N J Reusch 10/21/10

Author Signature Date