OPEN ACCESS

ORIGINAL ARTICLE

Comprehensive genomic analyses associate *UGT8* variants with musical ability in a Mongolian population

Hansoo Park,[1,2] Seungbok Lee,[1,3] Hyun-Jin Kim,[1,3] Young Seok Ju,[1,4] Jong-Yeon Shin,[1,5] Dongwan Hong,[1,6] Marcin von Grotthuss,[2] Dong-Sung Lee,[1,3] Changho Park,[7] Jennifer Hayeon Kim,[1] Boram Kim,[1] Yun Joo Yoo,[8] Sung-Il Cho,[9] Joohon Sung,[9] Charles Lee,[2] Jong-Il Kim,[1,3,5,7] Jeong-Sun Seo[1,3,4,5,7]

► Additional data are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2012-101209).

[1]Medical Research Center, Genomic Medicine Institute (GMI), Seoul National University, Seoul, Korea
[2]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[3]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea
[4]Macrogen Inc., Seoul, Korea
[5]Psoma Therapeutics Inc., Seoul, Korea
[6]Division of Convergence Technology, Functional Genomics Branch, National Cancer Center, Goyang, Korea
[7]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul, Korea
[8]Department of Mathematics Education, Seoul National University, Seoul, Korea
[9]Seoul National University School of Public Health, Seoul, Korea

Correspondence to
Dr Jeong-Sun Seo and Dr Jong-Il Kim, Genomic Medicine Institute, Medical Research Center, Seoul National University College of Medicine 101 Daehak-ro, Jongno-gu, Seoul 110-799, Korea;
jeongsun@snu.ac.kr, jongil@snu.ac.kr

HP, SL and H-JK contributed equally

Received 4 August 2012
Revised 25 September 2012
Accepted 10 October 2012

ABSTRACT

Background Musical abilities such as recognising music and singing performance serve as means for communication and are instruments in sexual selection. Specific regions of the brain have been found to be activated by musical stimuli, but these have rarely been extended to the discovery of genes and molecules associated with musical ability.

Methods A total of 1008 individuals from 73 families were enrolled and a pitch-production accuracy test was applied to determine musical ability. To identify genetic loci and variants that contribute to musical ability, we conducted family-based linkage and association analyses, and incorporated the results with data from exome sequencing and array comparative genomic hybridisation analyses.

Results We found significant evidence of linkage at 4q23 with the nearest marker D4S2986 (LOD=3.1), whose supporting interval overlaps a previous study in Finnish families, and identified an intergenic single nucleotide polymorphism (SNP) (rs1251078, $p=8.4\times10^{-17}$) near *UGT8*, a gene highly expressed in the central nervous system and known to act in brain organisation. In addition, a non-synonymous SNP in *UGT8* was revealed to be highly associated with musical ability (rs4148254, $p=8.0\times10^{-17}$), and a 6.2 kb copy number loss near *UGT8* showed a plausible association with musical ability ($p=2.9\times10^{-6}$).

Conclusions This study provides new insight into the genetics of musical ability, exemplifying a methodology to assign functional significance to synonymous and non-coding alleles by integrating multiple experimental methods.

INTRODUCTION

Song as a communication signal and as an instrument in sexual selection has been recognised since it was first proposed by Darwin.[1–3] Musical ability is a non-verbal and complex cognitive skill, and appears to have a latent biological basis in that infants can differentiate frequencies and 'carry a tune' without receiving extensive formal musical training.

Researchers have described certain aspects of how the architecture of the brain affects facets of musical ability. Perception and vocal production of singing seem to be based on the auditory and motor domains of the brain.[4 5] Studies of impaired language skills with spared musical abilities and impaired musical abilities with normal language skills have revealed a dissociation between these two skill sets,[6] leading to the proposal of a distinct mental module associated with separate neural substrates and a set of neurally isolatable processing components. A minority of humans exhibit extreme musical abilities in the form of either absolute pitch (the ability to accurately label tones with specific musical notes) or amusia (the inability to accurately identify and mimic tones).[7 8]

Recent studies have identified genetic components of musical ability. For example, absolute pitch has a significant familial basis and is predominant in females.[9] A twin study has shown substantial heritability for musical ability[10] and linkage studies have found loci for musical aptitude and absolute pitch.[11 12] Some polymorphisms of specific genes in association with musical ability have begun to be reported, including variants of *AVPR1A* and *SLC6A4*.[13 14]

As part of the GENDISCAN study (GENe DIScovery for Complex traits in large isolated families of Asians of the Northeast), which was designed to investigate genetic influences on complex traits in extended Asian families of rural Mongolia, we investigated the processing of pitch using 1008 subjects from 73 families. It was expected that several points of the GENDISCAN study would increase the power of genetic loci discovery in normal complex traits, considering (1) the study population has little ethnic admixture, (2) consists of large extended families, and (3) represents a community-based population unbiased by health status.[15]

To overcome the difficulties of identifying genetic variations underlying common complex diseases, an approach that allows for recruitment of homogeneous and isolated populations was proposed. However, only a few studies have incorporated this approach due to difficulties in sample recruitments. The inner Mongolian steppes are still inhabited by small populations; geographically isolated populations are commonly found in rural provinces of Mongolia. We recruited Mongolian individuals from an isolated population with large extended pedigrees. These individuals possess a

homogeneous genetic background and close genetic affinity to populations of the northern part of East Asia.[16–19]

Previously, binary familiarity tests have mostly been used to indicate whether or not each song part sounds similar to assess musical ability.[10 20–22] By shifting the pitch of melody one semitone higher or lower, participants were asked to classify two melodies as the same or different. In this study, we created a test to analyse subjects' acoustic outputs followed by hearing specific tones using cochlear implants (CI).[23 24] There are advantages to this approach, which include the possibility to study musical ability as a whole and the better availability of subjects. We determined the pitch discrimination limen with a simulated CI coding strategy and employed the complementary nature of linkage- and association-based methods for musical ability. The functional importance of results was screened through the incorporation of data from exome sequencing and array-based comparative genomic hybridisation (aCGH). This combined approach provides a method by which to discover additional novel genetic loci underlying complex traits.

METHODS

Study subjects and phenotype measurement

In 2006, a total of 2008 volunteers were recruited in Dashbalbar, Dornod Province, Mongolia for the GENDISCAN project,[25–28] which was designed to discover the genetic backgrounds of several complex traits (figure 1). For this project, we selected an isolated population composed of large extended families. This population is highly appropriate for gene mapping research due to its genetic homogeneity, decreased environmental heterogeneity, and restricted geographical distribution.[29] Extended multi-generation families comprising a small number of founders are known to increase the genetic power.[30] Traits included in this project are summarised in online supplementary table S1.

In this study, we chose 1008 individuals who are derived from 73 extended families and have precise pedigree structures. Table 1 lists descriptive characteristics of the study population. The average age of the participants is 31.0 years and 51.6% are women. The family structure in this population is very complicated, with multiple generations and many family pairs such as 1794 parent–offspring pairs, 734 full-siblings, 395 half-siblings, and 888 avuncular pairs. The average family size and standard deviation are 19.6 and 11.3, respectively. Peripheral blood sample was collected for each study subject, and DNA was extracted according to standard protocols. The extracted DNA was stored in solution at −20°C.

To examine the musical ability of subjects, we used a pitch-production accuracy (PPA) test based on the difference limen of a pitch paradigm in a psychophysical experiment with a simulated CI coding strategy.[31] PPA is given by $(100-10\times(|\nu_i-\nu_s|/\nu_s\times100))$, subtracting 10 points for each 1% error, where ν_s is the standard auditory frequency emitted by a pitch-producing device and ν_i is the vocal pitch frequency produced by the individuals, who hear a specific tone through a headset and recite the sound.[32] A harmonic tone complex with a sound pressure level of 70 dB intensity and sex-dependent fundamental frequency was used as a stimulus (see online supplementary table S2).

The participants with PPA values higher than 60 were categorised as individuals with good musical ability because they were consistently and accurately able to produce tones differing by less than a semitone from one another; the number of subjects with a PPA score over 60 was 357 (35.4%). However, for further analyses, participants with borderline PPA values between 50 and 70 were excluded to eliminate ambiguous PPA values; the number of subjects with PPA score over 70 was 268 (31.1%).

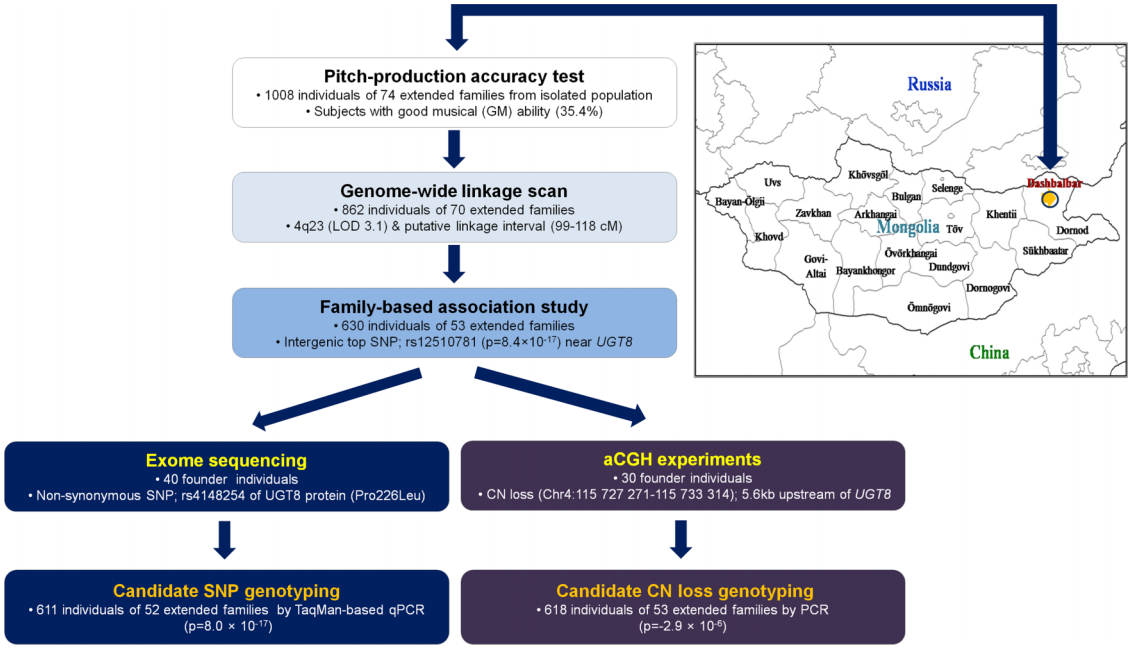


Figure 1 Overview of the project for musical ability. The pitch-production accuracy test was used to measure musical ability of 1008 individuals from 73 extended families of an isolated Mongolian population. We started with a genome-wide linkage study to identify potential causal loci associated with musical ability, and subsequently conducted a family-based association test under the linkage peak on 4q23 (99–118 cM). Furthermore, we used exome sequencing data in 40 founders and assessed copy number variants in 30 founders to explore plausible candidates for causal variants of musical ability with additional validating experiments. CN, copy number; SNP, single nucleotide polymorphism.

Table 1 Descriptive characteristics of study participants

Characteristics	Value
Sample information	
No. of samples	1008
No. of females (%)	520 (51.6)
Mean (SD) age (in years)	31.0 (15.5)
No. of sample with PPA score (%)	
≥70	268 (26.6)
≥60	357 (35.4)
<60	651 (64.6)
<50	594 (58.9)
Family information	
No. of families	73
Mean size (SD) of family members	19.61 (11.3)
No. of pairs	
Parent–offspring	1794
Full-sibling	734
Sister–sister	198
Brother–brother	167
Sister–brother	369
Half-sibling	395
Grandparent–grandchild	1202
Avuncular pairs	888
First cousins	598

PPA, pitch-production accuracy.

Genome-wide linkage scan and family-based association study under linkage region

We genotyped 862 samples from 70 families with deCODE 1039 microsatellite marker platform throughout the autosomes for genome-wide linkage analysis. We checked family relationships through PREST[33] using an average identity-by-descent (IBD)-based method. PEDCHECK was used to examine Mendelian inconsistencies in genotype data,[34] and non-Mendelian genotype errors were detected with SimWalk.[35] After fixing the genotype errors, multipoint identity-by-descent-matrices were calculated at each 1 cM distance, and converted using the Markov chain–Monte Carlo method by LOKI.[36] We used the Kosambi mapping function (derived from the deCODE map) to convert map distances into recombination fractions. For the multipoint linkage analyses, the Sequential Oligogenic Linkage Analysis Routines package was used.[37] We performed 10 000 permutation tests using the *lodadj* option to obtain the empirical p value. In addition, we estimated the adjusted narrow-sense heritability (h^2) (ie, the proportion of phenotype variance attributable to additive genetic variance). In all analyses, we used age and sex as covariates.

For further association analysis, 53 extended families composed of 630 family members were genotyped using an Illumina Human610-Quad BeadChip kit by Macrogen (Macrogen Inc, Seoul, Korea). We evaluated the Mendelian inconsistencies in single nucleotide polymorphism (SNP) data using PEDCHECK.[34] Non-Mendelian genotype errors were detected using Merlin.[38] SNP quality control assessment was based on SNP call rate, marker error rate, and minor allele frequency (MAF); minimum per-SNP call rate of 99%, less than 1% marker error rate, and higher than 5% MAF. In addition, we also removed genotypes with Hardy-Weinberg equilibrium p values $<1\times10^{-6}$. We focused on the putative linkage region in chromosome 4 for this analysis (1-LOD Unit Support Interval: 99–118 cM). A total of 3424 SNPs that met quality control criteria were included in the putative linkage region, and the PBAT tool in HelixTree software (V.6.4; GoldenHelix) was used for family-based association test (FBAT), which can control population stratification or population admixture.[15 39] The null hypothesis was 'linkage and no association (sandwich variance)',[40] which can be useful for expanded pedigrees by calculating a robust variance. We used the generalised estimating equation for the FBAT test statistic, and hypothesised an additive model. The association result was adjusted by covariates of age and sex.

Screening functional significance of candidates using exome sequencing and aCGH data integration

To assign a functional significance to candidates, we used exome sequencing data of 40 founders and 180K aCGH results of 30 founders, both of which were included in this study and previously genotyped in our group. The experimental summary of each is described in data supplement (see online supplementary tables S3–S5, supplementary methods). Among SNPs and short insertions/deletions (indels) called from exomes, we selected coding sequence SNPs and indels, and canonical splice-site variants as candidates, along with the copy number variants (CNVs) called from the aCGH experiment. Focusing on variants in the putative linkage region, we further narrowed our candidates by linkage disequilibrium (LD) estimation with the top 10 SNPs of our association study. Haploview software (V.3.2) was used for this LD estimation.

Among the candidates showing a significant level of LD, we selected one SNP and one CNV to be genotyped in our study population and compared their p values with the association results. For the SNP selected, three-dimensional (3D) modelling was conducted to predict its functional impact on the corresponding protein (see online supplementary methods).

RESULTS

Family-based linkage and association study

The heritability explained by the additive genetic portion of musical ability was estimated as 40% ($p<0.0001$, 95% CI 20.4% to 59.6%), and linkage regions with LOD>1.0 were found for musical ability from the genome-wide linkage scan (see online supplementary table S6). The maximum LOD score was 3.1 at chromosome 4q23 with the nearest marker D4S2986 (figure 2A), and the putative linkage region encompassing a maximum 1-LOD unit supports an interval range from 99 cM to 118 cM (figure 2B). In the next phase, we conducted FBAT to identify candidate variants within the putative linkage interval. Table 2 shows the top 10 SNPs that were significantly associated with musical ability, and all of these have reached the strict genome-wide significance of $p<1\times10^{-8}$. The strongest association ($p=8.4\times10^{-17}$) was found for rs12510781, an intergenic SNP near *UGT8* (MIM 601291). The regional association plot near *UGT8* is shown in figure 2C, and plotted recombination rates reflecting local LD structure were estimated from HapMap data. Three other SNPs (rs10024217, rs1903364, and rs12504058) were in moderate LD with rs12510781 ($r^2=0.4$). A synonymous SNP within *UGT8* (rs4148255) also showed significance in p value levels, despite the low LD with rs12510781 ($p=2.7\times10^{-10}$, $r^2<0.1$). The SNP with the second highest significance ($p=3.0\times10^{-13}$) was rs9307160 in the intron of *UNC5C* (MIM 603610), and the others were located near *ALPK1* (MIM 607347) and *ELOVL6* (MIM 611546).

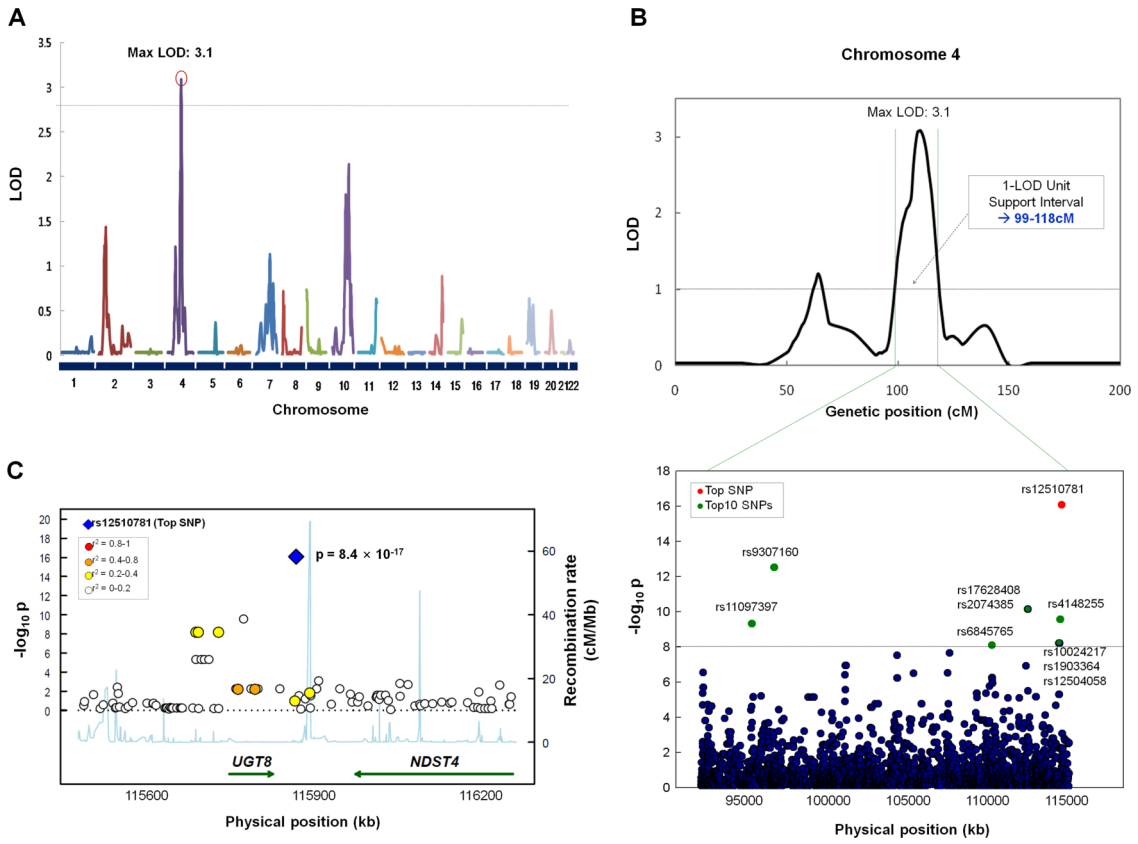


Figure 2 Summary of genome-wide linkage and association results for musical ability. (A) Genome-wide linkage results for musical ability. (B) The linkage peak on chromosome 4 and association plot under the linkage support region. The linkage support interval is indicated by a green line (99–118 cM). The red dot is the top single nucleotide polymorphism (SNP) by family-based association test. The SNPs 2–10 are labelled with green dots. (C) Regional plot of association results for SNPs from analysis ($-\log_{10}$ p) for *UGT8* (±300 kb position from top SNP). The SNPs close to rs12510781, the most significant SNP (blue diamond), are colour-coded to reflect their linkage disequilibrium with this SNP ($r^2<0.2$; white, $0.2\leq r^2<0.4$; yellow, $0.4\leq r^2<0.8$; orange, $r^2\geq 0.8$; red).

Utilisation of exome sequencing and aCGH data to assign functional significance to candidate variants

Among the candidates from the exome data (347 SNPs and seven indels in the putative linkage region), we narrowed down to four SNPs that were in strong LD with the top 10 SNPs identified via FBAT ($r^2>0.6$, online supplementary table S7). We found that a non-synonymous SNP (nsSNP) in *UGT8* (rs4148254) showed perfect LD with rs12510781, the most significant SNP from FBAT ($r^2=1.0$), and this SNP was genotyped in 611 FBAT samples for the association analysis. As a result, the LD between rs4148254 and rs12510781 was re-estimated ($r^2=0.93$), and the rs4148254 SNP was found to have the most significant association with musical ability in this study ($p=8.0\times10^{-17}$). The effect estimate of this SNP in founder samples was also higher than that of rs12510781 (OR=3.4, 95% CI 1.2 to 9.9 vs OR=3.0, 95% CI 1.1 to 8.2, online supplementary tables S8,S9). The 3D modelling of UGT8 protein showed that Pro226, which is changed to leucine by the SNP,

Table 2 Top 10 SNPs significantly associated with musical ability by FBAT under the putative linkage region of chromosome 4

SNP	*Position	Alleles		Frequency of effect allele	p Value (FBAT)	†Nearest gene(s)	Location (distance)
		Effect	Other				
rs12510781	115 860 030	G	A	0.12	8.4×10^{-17}	*UGT8*	Intergenic (42.3 kb)
rs9307160	96 586 977	C	T	0.10	3.0×10^{-13}	*UNC5C*	Intronic (−)
rs17628408	113 574 860	G	A	0.91	7.1×10^{-11}	*ALPK1*	Intronic (−)
rs2074385	113 598 098	C	A	0.91	7.1×10^{-11}	*ALPK1*	Intergenic (14.8 kb)
rs4148255	115 764 226	A	G	0.88	2.7×10^{-10}	*UGT8*	Synonymous (−)
rs11097397	95 087 875	G	T	0.28	4.8×10^{-10}	–	Intergenic (−)
rs10024217	115 677 564	C	T	0.28	6.1×10^{-9}	*UGT8*	Intergenic (61.4 kb)
rs1903364	115 681 713	C	T	0.28	6.1×10^{-9}	*UGT8*	Intergenic (57.3 kb)
rs12504058	115 718 566	G	A	0.28	6.1×10^{-9}	*UGT8*	Intergenic (20.4 kb)
rs6845765	111 177 613	C	T	0.86	8.2×10^{-9}	*ELOVL6*	Intergenic (12.0 kb)

*Positions are based on Build 36 from NCBI.
†Nearest gene, within ±100 kb of the SNP.
FBAT, family-based association test; SNP, single nucleotide polymorphism.

might be part of the loop exposed outside of the predicted 3D structure, and the loop with the Pro226 residue contains sequence motifs including TRFH domain docking and USP7-binding motifs (see online supplementary figure S1).

At the level of CNVs, only one copy number (CN) loss was found to have moderate LD with rs4148255, the fifth most significant SNP in FBAT (r^2=0.48; online supplementary table S10). This CN loss (Chr4: 115 727 257–115 733 452) is located 5.6 kb upstream of the *UGT8* gene. We genotyped it in 618 FBAT samples and the frequencies of heterozygous and homozygous CN losses were shown to be 45.15% and 10.03% in our study subjects (allele frequency=32.61%). This CNV was negatively associated with musical ability ($p=2.9\times10^{-6}$) and, interestingly, a diploid status at this position was shown to potentiate the positive effect of rs4148254 in founders (see online supplementary table S11). In addition, we identified a significant interaction effect between this CNV and rs4148254 using a logistic regression model (p=0.01).

DISCUSSION

In this study, we explored the genetic determinants of musical ability by combining several methodologies, namely family-based linkage and association studies supported by exome sequencing and aCGH data analyses. This study was conducted as a part of the GENDISCAN project, which was designed to discover the genetic backgrounds of complex traits in Mongolia.

Musical ability is a well-known complex trait determined by multiple environmental and genetic factors. As this trait consists of several factors including perception, cognition, learning, and emotions, a variety of genes have an effect on one's musical ability, both independently and interactively. To discover genetic backgrounds of these complex traits, studies should be designed from the first to increase the power to detect genetic loci. In this regard, our study has some strong points as described in the Introduction and Methods, which include little ethnic admixture and large extended families. In addition, we excluded samples with borderline phenotypes from all the analyses to derive more accurate results.

Our results support the view that musical ability is heritable and have shown significant evidence of linkage for musical ability in large families. Previously, a linkage study for musical aptitude was performed with samples in a small number of Finnish multigenerational families, composed of predominantly white subjects. That study found an association of the chromosomal region 4q22 with musical aptitude in the Finnish study population,[11] which overlaps with our linkage interval on chromosome 4q. Despite several differences in methodology, we believe that overlapping results for musical ability in different ethnic populations enhance the reliability of this linkage region on chromosome 4q.

We also discovered common variants strongly associated with musical ability, suggesting a biological mechanism for this finding. Including the most significant, five SNPs among the top 10 were shown to lie near or within *UGT8*. In addition, there was no LD structure between rs12510781 and rs4148255. These two unrelated variants on one gene, associated with the same phenotype, increase the possibility of *UGT8* being one of the true susceptibility genes for musical ability.

To identify more detailed causal variants, we integrated additional technologies such as exome sequencing and aCGH, resulting in the discovery of another nsSNP in *UGT8* and a CN loss located 5.6 kb upstream of this gene. The SNP rs4148254, which changes amino acid 226 of the UGT8 protein from proline to leucine, was not included in the platform we used, and has shown a lower p value than rs12510781 in our study population (see online supplementary figure S1A,B). Because the BLOSUM score[41] for this change is '–3', and PolyPhen-2[42] predicts this to be damaging, the SNP might affect the function of the UGT8 protein. Moreover, this proline amino acid seems to be conserved among vertebrates (see online supplementary table S12). The three other SNPs (rs35308602, rs2074381, and rs3828539), which were in high LD (r^2>0.6) with the top 10 SNPs, were predicted to be benign by PolyPhen-2 and the BLOSUM scores were '2', '1', and '–1', respectively (see online supplementary table S7). In case of the CN loss, even though it was not more significant than the associated SNP allele, the synergetic effect of this variant with rs4148254 was suggested in the founder analysis.

The protein encoded by *UGT8* is UDP glycosyltransferase 8, which is highly expressed in brain (see online supplementary figure S2). It is the first enzyme involved in complex lipid biosynthesis in the myelinating oligodendrocyte[43] and clearance of long-chain ceramides (lcCer). lcCer clearance in neurons is mediated by glucosylceramide synthase (GCS) and studies have shown that decreased GCS leads to abnormally high lcCer.[44] A significant early downregulation in glial GCS expression was associated with an increase in *UGT8* mRNA in Alzheimer's disease,[45] and some patients with Alzheimer's disease have been observed to preserve musical ability long after losing all other cognitive functions.[6]

Although this study primarily focused on *UGT8*, there are other genes such as *UNC5C*, *ALPK1*, and *ELOVL6* equally worth our attention. The protein encoded by *UNC5C* plays a role in the chemorepulsive effect of netrin-1 in axon guidance. This gene was previously suggested as a susceptibility gene for musical ability in the Finnish linkage study.[11] Regarding the other two, one study has shown that mice homozygous for disrupted copies of Alpk1 exhibited coordination defects,[46] and *ELOVL6* was once reported as one of the susceptibility loci for attention-deficit/hyperactivity disorder in a genome-wide association study.[47] Several previous findings, as listed above, have supported the neural involvement of those candidate genes; however, more evidence should be given to associate them with musical ability.

Music is a complex cognitive skill in the neuronal network affected by several potential covariates. We first considered language ability as a potential covariate besides age and sex. However, we found no language skill defects in our study subjects, and previous studies have reported that it is possible for language skills to be impaired while musical abilities are spared (aphasia without amusia); likewise, musical abilities can be impaired while language skills are spared (amusia without aphasia).[6 48] In addition, more factors including special musical training, education status, and education duration might be considered as potential covariates, since it has been reported that the skill of absolute pitch could be developed at a very young age by special musical training.[49 50] However, our participants lived in an isolated area with a homogeneous culture, and most of them were educated in the same public school without any additional musical training. In this study, therefore, we did not take those factors into account for analyses.

In summary, we have demonstrated for the first time that common genetic variants in *UGT8* are associated with musical ability, exemplifying a methodology to assign functional significance to the results of various association studies, which in many cases yield synonymous or non-coding alleles.

Acknowledgements The authors appreciate the help of all study participants and collaborators. We thank Omer Gokcumen and Raju Govindaraju at Harvard Medical School and Thomas Bleazard at Seoul National University for their personal comments regarding this manuscript.

Contributors J-SS planned and managed the project. J-IK, HP, YSJ, S-IC, and JS recruited and measured phenotypes of the Mongolian samples. HP, H-JK, YJY and J-IK analysed linkage data and family-based association studies. SL, J-YS, DH, and J-IK executed exome sequencing and analysed sequence data. HP, SL, D-SL, CP, JHK, and BK executed and analysed aCGH experiments. MvG performed the 3D modelling and motif analysis. CL supervised research at Brigham and Women's Hospital/Harvard Medical School. J-SS, HP, H-JK, SL, and J-IK wrote the manuscript and CL edited the manuscript.

Funding This work was supported by the Korean Ministry of Education, Science and Technology (Grant No. 2003-2001558) and the US National Institutes of Health (Grant No. HG004221).

Competing interests None.

Ethics approval This study was approved by the Institutional Review Board of the Seoul National University Hospital (approval number, H-0307-105-002).

Data sharing statement The whole data of exome and aCGH experiments, some of which were used for this study, have not been published yet. We can provide the part of data related to this project upon request.

REFERENCES

1. **Darwin C.** *The descent of man, and selection in relation to sex*. New York: D. Appleton and company, 1871.
2. **Miller GF.** *The mating mind: how sexual choice shaped the evolution of human nature*. 1st edn. New York: Doubleday, 2000.
3. **Fitch WT.** The biology and evolution of music: a comparative perspective. *Cognition* 2006;**100**:173–215.
4. **Peretz I,** Coltheart M. Modularity of music processing. *Nat Neurosci* 2003;**6**:688–91.
5. **Coltheart M.** Modularity and cognition. *Trends Cogn Sci* 1999;**3**:115–20.
6. **Halpern AR,** O'Connor MG. Implicit memory for music in Alzheimer's disease. *Neuropsychology* 2000;**14**:391–7.
7. **Peretz I,** Cummings S, Dube MP. The genetics of congenital amusia (tone deafness): a family-aggregation study. *Am J Hum Genet* 2007;**81**:582–8.
8. **Peretz I,** Ayotte J, Zatorre RJ, Mehler J, Ahad P, Penhune VB, Jutras B. Congenital amusia: a disorder of fine-grained pitch discrimination. *Neuron* 2002;**33**:185–91.
9. **Profita J,** Bidder TG. Perfect pitch. *Am J Med Genet* 1988;**29**:763–71.
10. **Drayna D,** Manichaikul A, de Lange M, Snieder H, Spector T. Genetic correlates of musical pitch recognition in humans. *Science* 2001;**291**:1969–72.
11. **Pulli K,** Karma K, Norio R, Sistonen P, Goring HH, Jarvela I. Genome-wide linkage scan for loci of musical aptitude in Finnish families: evidence for a major locus at 4q22. *J Med Genet* 2008;**45**:451–6.
12. **Theusch E,** Basu A, Gitschier J. Genome-wide study of families with absolute pitch reveals linkage to 8q24.21 and locus heterogeneity. *Am J Hum Genet* 2009;**85**:112–19.
13. **Ukkola LT,** Onkamo P, Raijas P, Karma K, Jarvela I. Musical aptitude is associated with AVPR1A-haplotypes. *PLoS One* 2009;**4**:e5534.
14. **Morley AP,** Narayanan M, Mines R, Molokhia A, Baxter S, Craig G, Lewis CM, Craig I. AVPR1A and SLC6A4 polymorphisms in choral singers and non-musicians: a gene association study. *PLoS One* 2012;**7**:e31763.
15. **Ott J,** Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nat Rev Genet* 2011;**12**:465–74.
16. **Katoh T,** Mano S, Ikuta T, Munkhbat B, Tounai K, Ando H, Munkhtuvshin N, Imanishi T, Inoko H, Tamiya G. Genetic isolates in East Asia: a study of linkage disequilibrium in the X chromosome. *Am J Hum Genet* 2002;**71**:395–400.
17. **Nakajima T,** Wooding S, Sakagami T, Emi M, Tokunaga K, Tamiya G, Ishigami T, Umemura S, Munkhbat B, Jin F, Guan-Jun J, Hayasaka I, Ishida T, Saitou N, Pavelka K, Lalouel JM, Jorde LB, Inoue I. Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. *Am J Hum Genet* 2004;**74**:898–916.
18. **Katoh T,** Munkhbat B, Tounai K, Mano S, Ando H, Oyungerel G, Chae GT, Han H, Jia GJ, Tokunaga K, Munkhtuvshin N, Tamiya G, Inoko H. Genetic features of Mongolian ethnic groups revealed by Y-chromosomal analysis. *Gene* 2005;**346**:63–70.
19. **Kimura T,** Kobayashi T, Munkhbat B, Oyungerel G, Bilegtsaikhan T, Anar D, Jambaldorj J, Munkhsaikhan S, Munkhtuvshin N, Hayashi H, Oka A, Inoue I, Inoko H. Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33–q23 in Mongolians. *Hum Genet* 2008;**123**:655–60.
20. **Kalmus H,** Fry DB. On tune deafness (dysmelodia): frequency, development, genetics and musical background. *Ann Hum Genet* 1980;**43**:369–82.
21. **Peretz I,** Gaudreau D, Bonnel AM. Exposure effects on music preference and recognition. *Mem Cognit* 1998;**26**:884–902.
22. **Ayotte J,** Peretz I, Hyde K. Congenital amusia: a group study of adults afflicted with a music-specific disorder. *Brain* 2002;**125**(Pt 2):238–51.
23. **Leal MC,** Shin YJ, Laborde ML, Calmels MN, Verges S, Lugardon S, Andrieu S, Deguine O, Fraysse B. Music perception in adult cochlear implant recipients. *Acta Otolaryngol* 2003;**123**:826–35.
24. **Haumann S,** Mühler R, Ziese M, Specht H. Diskrimination musikalischer Tonhöhen bei Patienten mit Kochleaimplantat. *HNO* 2007;**55**:613–19.
25. **Ju YS,** Park H, Lee MK, Kim JI, Sung J, Cho SI, Seo JS. A genome-wide Asian genetic map and ethnic comparison: the GENDISCAN study. *BMC Genomics* 2008;**9**:554.
26. **Im SW,** Kim HJ, Lee MK, Yi JH, Jargal G, Sung J, Cho SI, Kim JI. Genome-wide linkage analysis for ocular and nasal anthropometric traits in a Mongolian population. *Exp Mol Med* 2010;**42**:799–804.
27. **Paik SH,** Kim HJ, Son HY, Lee S, Im SW, Ju YS, Yeon JH, Jo SJ, Eun HC, Seo JS, Kwon OS, Kim JI. Gene mapping study for constitutive skin color in an isolated Mongolian population. *Exp Mol Med* 2012;**44**:241–9.
28. **Lee MK,** Cho SI, Kim H, Song YM, Lee K, Kim JI, Kim DM, Chung TY, Kim YS, Seo JS, Ham DI, Sung J. Epidemiologic characteristics of intraocular pressure in the Korean and Mongolian populations: the Healthy Twin and the GENDISCAN study. *Ophthalmology* 2012;**119**:450–7.
29. **Kristiansson K,** Naukkarinen J, Peltonen L. Isolated populations and complex disease gene identification. *Genome Biol* 2008;**9**:109.
30. **Arcos-Burgos M,** Muenke M. Genetics of population isolates. *Clin Genet* 2002;**61**:233–47.
31. **Haumann S,** Muhler R, Ziese M, von Specht H. (Discrimination of musical pitch with cochlear implants). *HNO* 2007;**55**:613–19.
32. **Nikjeh DA,** Lister JJ, Frisch SA. The relationship between pitch discrimination and vocal production: comparison of vocal and instrumental musicians. *J Acoust Soc Am* 2009;**125**:328–38.
33. **McPeek MS,** Sun L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* 2000;**66**:1076–94.
34. **O'Connell JR,** Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998;**63**:259–66.
35. **Weeks DE,** Sobel E, O'Connell JR, Lange K. Computer programs for multilocus haplotyping of general pedigrees. *Am J Hum Genet* 1995;**56**:1506–7.
36. **Heath SC.** Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997;**61**:748–60.
37. **Almasy L,** Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998;**62**:1198–211.
38. **Abecasis GR,** Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;**30**:97–101.
39. **Lange C,** DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: tools for family-based association studies. *Am J Hum Genet* 2004;**74**:367–9.
40. **Laird NM,** Lange C. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 2006;**7**:385–94.
41. **Henikoff S,** Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–19.
42. **Adzhubei IA,** Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
43. **Dugas JC,** Tai YC, Speed TP, Ngai J, Barres BA. Functional genomic analysis of oligodendrocyte differentiation. *J Neurosci* 2006;**26**:10967–83.
44. **Han X,** D MH, McKeel DW Jr, Kelley J, Morris JC. Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: potential role in disease pathogenesis. *J Neurochem* 2002;**82**:809–18.
45. **Marks N,** Berg MJ, Saito M. Glucosylceramide synthase decrease in frontal cortex of Alzheimer brain correlates with abnormal increase in endogenous ceramides: consequences to morphology and viability on enzyme suppression in cultured primary neurons. *Brain Res* 2008;**1191**:136–47.
46. **Chen M,** Xu R. Motor coordination deficits in Alpk1 mutant mice with the inserted piggyBac transposon. *BMC Neurosci* 2011;**12**:1.
47. **Mick E,** Todorov A, Smalley S, Hu X, Loo S, Todd RD, Biederman J, Byrne D, Dechairo B, Guiney A, McCracken J, McGough J, Nelson SF, Reiersen AM, Wilens TE, Wozniak J, Neale BM, Faraone SV. Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry* 2010;**49**:898–905 e3.
48. **Barlett JC,** Halpern AR, Dowling WJ. Recognition of familiar and unfamiliar melodies in normal aging and Alzheimer's disease. *Mem Cognit* 1995;**23**:531–46.
49. **Takeuchi AH,** Hulse SH. Absolute pitch. *Psychol Bull* 1993;**113**:345–61.
50. **Gregersen PK,** Kowalsky E, Kohn N, Marvin EW. Early childhood music education and predisposition to absolute pitch: teasing apart genes and environment. *Am J Med Genet* 2001;**98**:280–2.

Supplementary table 1 Traits measured in the GENDISCAN project

Categories	Traits
Anthropometric Measurement	Height, Weight, Sitting Height, Head Circumference, Waist Circumference, Hip Circumference, Thigh Length, Leg Length, Calf Length, Foot Length, Foot Width, Body Mass Index, Face
Cardiovascular Parameter	Systolic Blood Pressure, Diastolic Blood Pressure, Pulse Pressure, Mean Arterial Pressure, Heart Rate, Electrocardiogram
Ophthalmologic Parameter	Tonometry, Visual Accuracy, Schirmer Test, Refractometry
Pulmonary Function Test	Forced Expiratory Volume in 1 sec (FEV1), Forced Vital Capacity (FVC), FEV1 / FVC
Allergy / Atopy	Grass, Tree, Murgurt, Ragweet, Cockroach, Dog, Cat, Horse, Cow, Goat, Alternaria, Aspergillus
Whole Body Impedence	Fat%, Lean Body Mass, Total Body Water
Blood Test	Complete Blood Cell Count, Biochemistry Test, Electrolyte Test
Urine Test	Microalbumin, Creatinine, Microalbumin/Creatinine Ratio, Ca, Phosphate, Uric Acid
Others	Musical Ability, HOMA, Glomerular Filtration Rate, Bone Mineral Density, Skin Color, etc.

Supplementary table 2 Frequencies of musical notes for male and female

Male	C3	C#3	D3	D#3	E3	F3	F#3	G3	G#3	A3	A#3	B3	C4
(Hz)	(131)	(139)	(147)	(156)	(165)	(175)	(185)	(196)	(208)	(220)	(233)	(247)	(262)
Female	C4	C#4	D4	D#4	E4	F4	F#4	G4	G#4	A4	A#4	B4	C5
(Hz)	(262)	(277)	(294)	(311)	(330)	(349)	(370)	(392)	(415)	(440)	(466)	(494)	(523)

Supplementary table 3 Exome sequencing summary

Sample	Total Read	Aligned Read (Unique)	% Covered Bait Base (≥1x)	% Covered Bait Base (≥4x)	Mean Coverage Depth of Baits (x)
E_f1	32 852 908	27 106 827	86.0	72.4	20.2
E_f2	39 589 740	32 283 923	88.7	77.4	26.2
E_f3	41 213 346	33 649 519	89.4	78.8	27.8
E_f4	20 941 378	16 789 898	86.8	72.2	14.3
E_f5	29 748 524	24 339 510	90.4	79.8	25.0
E_f6	42 984 976	34 816 307	93.0	85.2	34.0
E_f7	36 877 332	29 961 171	92.2	83.5	29.9
E_f8	37 912 008	30 834 473	92.7	84.3	31.7
E_f9	36 366 364	30 015 971	89.4	78.5	25.3
E_f10	37 300 546	30 375 062	92.1	83.3	31.4
E_f11	42 188 960	34 422 535	92.8	84.7	33.4
E_f12	36 166 074	29 732 171	88.3	77.3	28.2
E_f13	37 237 784	30 542 925	89.5	79.0	28.1
E_f14	37 706 520	30 858 648	88.7	77.9	28.5
E_f15	37 685 392	31 053 639	88.8	77.9	28.8
E_f16	42 778 900	34 532 430	89.9	80.0	31.0
E_f17	35 344 498	29 236 251	88.3	77.4	27.1
E_f18	36 483 606	29 341 728	87.2	75.5	26.8
E_f19	34 093 870	27 626 902	87.0	74.9	25.3

E_f20	36 177 456	29 167 664	87.2	75.5	26.3
E_f21	41 208 568	33 552 315	87.9	76.7	30.2
E_f22	39 213 348	31 575 226	87.8	76.6	28.6
E_f23	19 852 910	15 856 982	85.5	71.1	15.4
E_f24	33 841 666	27 783 724	83.7	70.6	24.4
E_f25	42 573 026	34 961 147	86.2	73.7	28.6
E_f26	39 918 988	33 045 045	85.4	72.8	27.8
E_f27	38 038 236	31 244 566	84.9	72.1	26.6
E_f28	38 021 060	31 167 364	84.8	72.1	26.6
E_f29	32 781 766	27 012 922	84.4	70.8	23.1
E_f30	34 835 446	30 971 369	94.0	83.9	25.8
E_f31	38 594 842	31 358 956	85.4	72.3	22.9
E_f32	39 185 464	32 156 651	85.1	71.8	23.5
E_f33	39 015 288	32 446 516	84.4	71.2	23.8
E_f34	35 869 264	29 461 527	83.7	70.3	21.9
E_f35	34 262 894	27 855 979	86.1	72.7	20.9
E_f36	19 149 382	15 368 514	84.4	67.6	12.0
E_f37	39 723 174	32 743 894	87.9	75.8	24.7
E_f38	36 467 176	30 006 354	87.7	75.3	22.8
E_f39	43 120 662	38 658 825	95.4	87.5	32.1
E_f40	32 022 542	26 295 343	86.6	73.2	19.8

Supplementary table 4 SNP and indel summary of exome sequencing

Sample	SNPs			Indels		
	Total	*CDS (nsSNP)	Splicing Site	Total	CDS (frameshift)	Splicing Site
E_f1	71 500	11 713 (5 839)	42	4 281	145 (74)	11
E_f2	86 454	12 652 (6 211)	45	5 205	149 (70)	10
E_f3	88 008	13 222 (6 511)	41	5 328	164 (93)	12
E_f4	41 844	11 180 (5 552)	36	2 425	133 (65)	9
E_f5	77 249	13 698 (6 744)	47	4 439	170 (88)	7
E_f6	92 963	15 303 (7 388)	56	5 156	186 (97)	14
E_f7	77 066	14 536 (6 984)	49	4 438	172 (93)	11
E_f8	89 529	14 967 (7 243)	54	5 039	187 (94)	8
E_f9	73 755	12 907 (6 317)	49	4 354	155 (73)	10
E_f10	89 105	14 724 (7 132)	51	5 019	197 (108)	11
E_f11	86 969	15 006 (7 301)	47	4 904	175 (100)	8
E_f12	95 318	12 906 (6 393)	45	5 891	166 (93)	9
E_f13	96 013	13 571 (6 680)	50	5 651	170 (85)	11
E_f14	102 540	13 105 (6 458)	47	6 193	158 (85)	10
E_f15	94 552	13 211 (6 514)	50	5 566	168 (88)	11
E_f16	119 540	13 628 (6 689)	44	7 073	154 (80)	7
E_f17	86 307	13 201 (6 455)	43	5 009	151 (84)	6
E_f18	87 616	12 372 (6 080)	49	5 106	156 (92)	9
E_f19	69 364	12 244 (6 084)	40	4 295	155 (83)	6

E_f20	94 033	12 374 (6 127)	40	5 587	147 (80)	14
E_f21	73 301	12 518 (6 179)	44	4 753	173 (97)	10
E_f22	107 969	12 644 (6 211)	46	6 541	138 (76)	7
E_f23	40 162	10 876 (5 416)	36	2 346	121 (67)	9
E_f24	73 343	11 819 (6 359)	71	4 246	133 (72)	10
E_f25	93 098	11 886 (5 963)	51	5 647	136 (70)	9
E_f26	75 492	11 617 (5 737)	40	4 648	152 (83)	8
E_f27	86 691	11 815 (6 059)	58	5 120	132 (70)	9
E_f28	88 870	11 998 (6 284)	76	5 080	146 (83)	9
E_f29	71 241	11 521 (5 945)	53	4 471	136 (67)	6
E_f30	82 696	13 759 (6 606)	47	3 627	125 (51)	5
E_f31	107 913	12 728 (6 849)	76	6 046	156 (92)	8
E_f32	105 726	12 154 (6 302)	56	5 850	135 (76)	6
E_f33	87 682	12 331 (6 468)	75	5 156	133 (64)	14
E_f34	76 020	12 242 (6 666)	106	4 822	147 (82)	12
E_f35	83 440	12 035 (6 089)	49	4 800	147 (76)	12
E_f36	46 436	10 596 (5 435)	37	2 572	101 (56)	9
E_f37	78 871	12 549 (6 179)	43	5 075	152 (85)	9
E_f38	88 679	12 442 (6 271)	46	5 098	151 (77)	10
E_f39	87 191	14 716 (7 064)	67	4 352	126 (61)	11
E_f40	62 335	11 852 (5 996)	37	3 992	139 (72)	7

*CDS, coding sequence

Supplementary table 5 CNV summary of founder samples

Sample ID	Total Size of *CN Gains	Total Size of CN Losses	†# of CN Gains	# of CN Losses	# of CN gains of chr4	# of CN losses of chr4
C_f1	15 623 979	10 026 049	867	984	27	79
C_f2	11 034 602	9 619 206	489	994	28	47
C_f3	10 655 154	9 043 016	485	975	22	80
C_f4	11 721 523	8 302 210	440	979	25	53
C_f5	17 099 926	8 492 324	689	991	20	35
C_f6	10 510 945	8 798 506	447	983	20	58
C_f7	14 433 779	9 342 444	520	1,005	35	52
C_f8	12 586 640	9 498 884	542	998	27	76
C_f9	11 441 112	8 364 814	468	920	13	28
C_f10	10 309 943	8 485 796	435	862	12	23
C_f11	11 919 407	9 406 104	677	1,159	29	92
C_f12	11 319 754	7 881 799	524	915	16	26
C_f13	11 479 146	9 236 424	481	1,000	16	29
C_f14	10 531 552	8 249 997	373	887	21	20
C_f15	8 280 327	9 618 603	593	965	17	70
C_f16	13 270 274	11 033 706	726	1,223	35	91
C_f17	12 498 673	8 531 715	559	980	22	31
C_f18	11 319 929	9 661 462	576	959	24	68
C_f19	12 253 256	7 909 510	628	884	29	22
C_f20	10 830 936	8 449 469	417	958	21	51

C_f21	13 211 829	8 991 969	620	931	17	27
C_f22	10 759 939	10 101 000	489	1,036	21	76
C_f23	11 081 237	9 219 670	462	1,058	25	60
C_f24	11 152 189	12 678 057	793	986	26	54
C_f25	12 095 182	9 317 092	433	939	18	25
C_f26	14 948 682	8 703 857	807	1,174	30	88
C_f27	11 197 216	8 780 281	468	941	20	29
C_f28	11 496 936	9 594 463	482	1,004	18	65
C_f29	9 562 783	9 553 762	357	952	16	25
C_f30	11 245 642	8 674 725	422	980	20	60

*CN, copy number; †# , number

Supplementary table 6 Heritability and linkage regions from genome-wide linkage scan for musical ability

Chromosome (location, cM)	Maximum LOD score	Nearest marker	Locus	1-LOD unit support interval (cM)	Empirical p Value
2 (67)	1.4	D2S2328	2p22.1	54-71	0.0059
4 (110)	3.1	D4S2986	4q23	99-118	< 0.0001
10 (142)	2.1	D10S562	10q25.3	112-150	0.0011
Narrow sense heritability (%); h^2 (95% CI) = 40 (20.4 ~ 59.6), p<0.0001					

Supplementary table 7 LD estimation between 4 candidate SNPs identified by exome sequencing and top 10 SNPs of FBAT

Top 10 SNPs	4 candidates	r^2	Gene (amino acid change)	PolyPhen-2 prediction	BLOSUM score
rs12510781	rs4148254	1.0	*UGT8* (P226L)	probably damaging	-3
rs17628408	rs35308602	1.0	*ALPK1* (E910D)	benign	2
rs2074385		1.0			
rs17628408	rs2074381	1.0	*ALPK1* (N916D)	benign	1
rs2074385		1.0			
rs17628408	rs3828539	0.7	*C4orf21* (I232T)	benign	-1
rs2074385		0.7			

Supplementary table 8 The odds ratio of top SNP (rs12510781) allele in founder samples (n=103)

	Control	Good musical ability	Total	*OR (95% CI)	*†adjust OR (95% CI)	†p Value
A (other allele)	148	31	179	2.6 (1.0-6.9)	3.0 (1.1-8.2)	0.031
G (effect allele)	18	9	27			
Total	166	40	206			

*OR and 95% CI was estimated by logistic regression analysis under additive genetic model.

†OR and p Value were adjusted by covariates such as age and sex.

Supplementary table 9 The odd ratio of nsSNP (rs4148254) allele in founder samples (n=97)

	Control	Good musical ability	Total	*OR (95% CI)	*†adjust OR (95% CI)	†p Value
C (other allele)	142	30	172	2.9 (1.0-7.9)	3.41 (1.2-9.9)	0.024
T (effect allele)	14	8	22			
Total	156	38	194			

*OR and 95% CI was estimated by logistic regression analysis under additive genetic model.

†OR and p Value were adjusted by covariates such as age and sex.

Supplementary table 10 LD analysis between 10 candidate SNPs and CNVs within the support linkage interval on chromosome 4

CNV regions	Gain/Loss	top 10 SNPs	r^2
chr4:115727257-115733452	Loss	rs4148255	0.48
chr4:111602578-111603069	Loss	rs6845765	0.1
chr4:115615545-115617174	Gain	rs10024217	0.1
chr4:115615545-115617174	Gain	rs1903364	0.1
chr4:115615545-115617174	Gain	rs12504058	0.1
chr4:115963329-115963581	Loss	rs4148255	0.06
chr4:115963329-115963581	Gain	rs4148255	0.06
chr4:111190885-111192069	Loss	rs6845765	0.05
chr4:115391466-115403784	Loss	rs10024217	0.04
chr4:115391466-115403784	Loss	rs1903364	0.04
chr4:115391466-115403784	Loss	rs12504058	0.04
chr4:115615545-115617174	Gain	rs12510781	0.03
chr4:96379890-96380568	Gain	rs9307160	0.03
chr4:115727257-115733452	Loss	rs10024217	0.02
chr4:115727257-115733452	Loss	rs1903364	0.02
chr4:115727257-115733452	Loss	rs12504058	0.02

Supplementary table 11 The odd ratio of nsSNP (rs4148254) allele in founder samples without CN loss (Chr4:115 727 257-115 733 452) (n=36)

	Control	Good musical ability	Total	*OR (95% CI)	*†adjust OR (95% CI)	†p Value
C (other allele)	55	10	65	5.93 (1.1-33.0)	11.7 (1.2-116.5)	0.037
T (effect allele)	3	4	7			
Total	58	14	72			

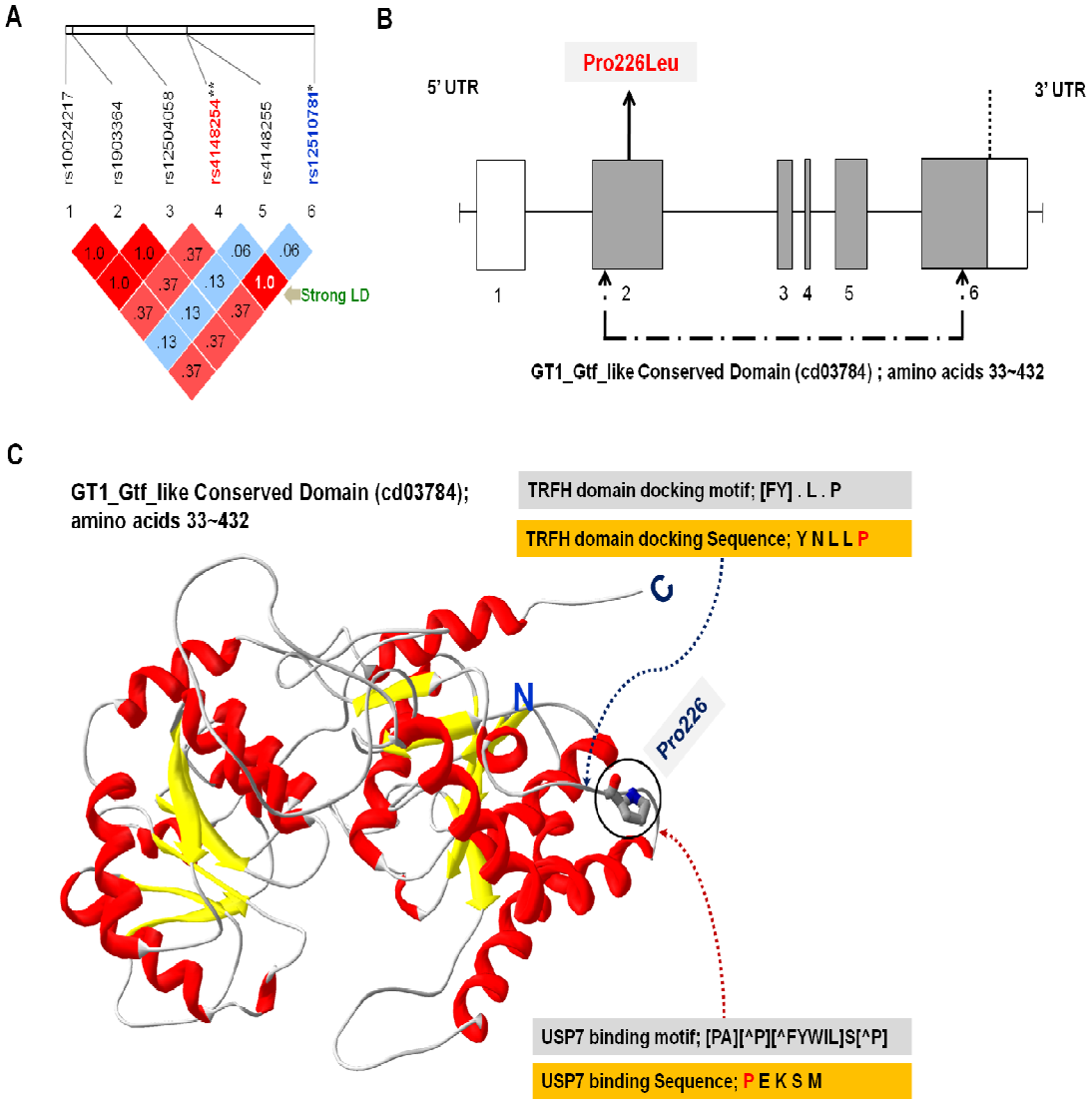
*OR and 95% CI was estimated by logistic regression analysis under additive genetic model.

†OR and p Value were adjusted by covariates such as age and sex.

Supplementary table 12 Amino acid 226 of UGT8 protein conserved among vertebrate species (HomoloGene release 64)

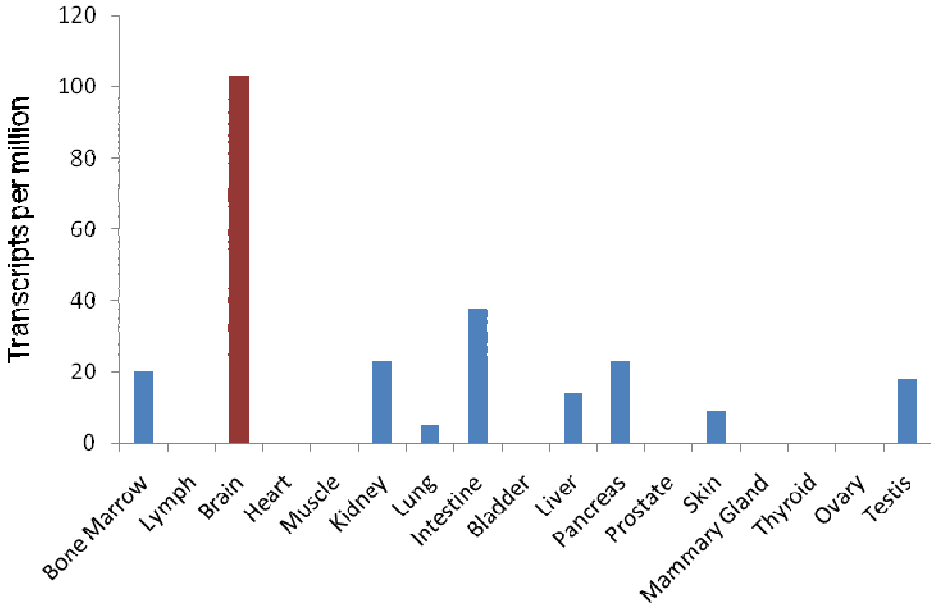
Species	Protein ID	Amino acid sequence
Homo sapiens	NP_001121646.1	217 R I M Q K Y N L L **P** E K S M Y D L V H G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Pan troglodytes	XP_001146745.1	217 R I M Q K Y N L L **P** E K S M Y D L V H G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Canis lupus familiaris	XP_545033.2	217 R I M Q K Y N L L **P** E K S M Y D L V H G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Bos taurus	NP_001077104.1	217 R I M Q K Y N L L **P** E K S M Y D L V Y G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Mus musculus	NP_035804.2	217 R I M Q K Y N L L **P** A K S M Y D L V H G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Rattus norvegicus	NP_062149.1	217 R I M Q K Y N L L **P** A K S M Y D L V H G S - S L W M L C T D V A L E F P R P T L P N V V Y V G G I - 264
Gallus gallus	NP_989535.1	217 R I M Q K H K V L **P** E R S M Y D L V H G S - S L W M L C T D I A L E F P R P T L P N V V Y V G G I - 264
Danio rerio	NP_001037790.1	218 R I M R K Y N I Q **P** S V S M H D L V Q N S - R L W M L C T D M A L E F P R P T L P H V V Y V G G I - 265

Supplementary figure 1

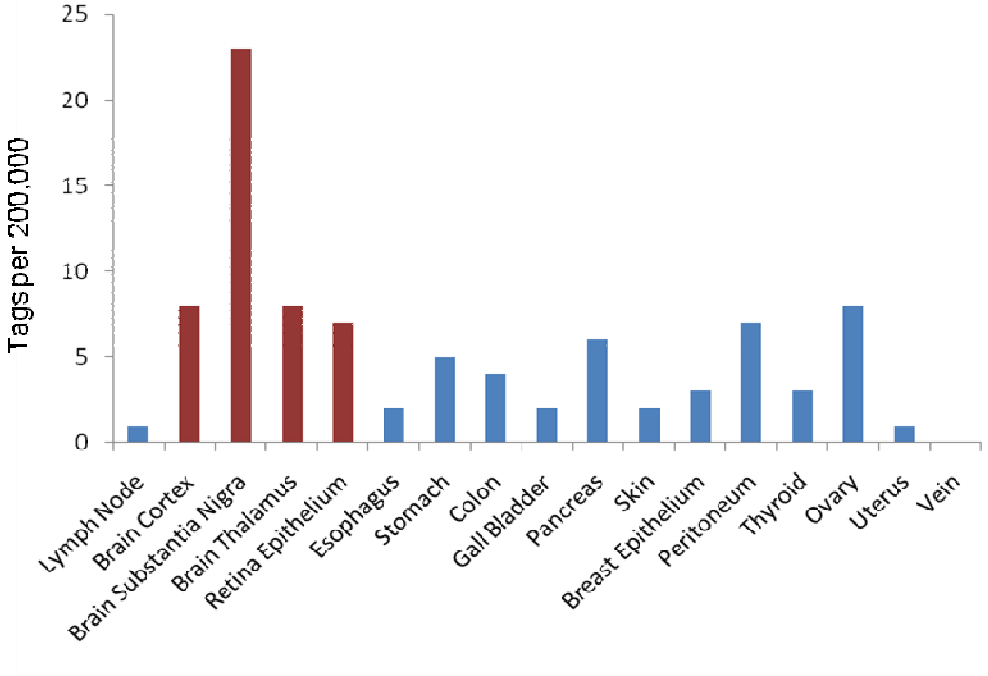


Supplementary figure 2

(A)



(B)



SUPPLEMENTARY METHODS

Exome sequencing

40 founder individuals were sequenced through exome sequencing. After bar-coding with Multiplexing Sample Preparation Oligonucleotide Kit (Illumina, Inc.), the samples were captured with the SureSelect Human All Exon Kit (Version 1.0.1, Agilent, Inc.). Each captured DNA was fragmented into 200-250 bp pieces, which were paired-end sequenced using an Illumina Genome Analyzer IIx. The read length for this experiment was uniformly 72 bp. The generated reads were aligned to human reference NCBI build 36.3 with GSNAP alignment tool,[1] and base variations including single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) were called. SNP calling criteria were as follows; 1) No less than 4 uniquely aligned reads indicate a variant, 2) Including random alignment, at least 20% of the reads must agree for heterozygous variants and at least 90% must agree for homozygous variants, 3) Mean quality score (Q score) for variant is no less than 20. Regarding indel calling, the only difference in criteria is the proportion of reads required; a proportion between 20% and 60% is determined as a heterozygote, and a proportion higher than 60% is a homozygote.[2] Gene annotation for variants was based on the RefSeq gene set.

180k probes aCGH array

The copy number variation (CNV) targeted custom array CGH (aCGH) platform was manufactured in 4x180k format on SurePrint G3 Human CGH Microarrays (Agilent, Inc.). This format provides more than 180 000 probes on one quarter of a glass microscope slide and allows for the interrogation of thousands of known CNVs simultaneously in a single sample. We used the set of 8 599 CNVs that were

identified by the Structural Genomic Variation Consortium. Next, we included regions of the 4 317 deletions released in June 2009 as part of the 1000 Genomes Project. Third, we incorporated 3 547 Asian specific CNVs discovered by a high resolution 24M feature probe set.[3] In addition, a set of known segmental duplications and novel sequences identified in the HuRef genome, and the regions catalogued in Database of Genomic Variants (http://projects.tcag.ca/variation/) that do not overlap with the above mentioned datasets were included. These probes are then assessed in various ways, including median Log2Ratio, median r-channel (red) signal intensity, and median g-channel (green) signal intensity. Additionally, the sequence coverage of NA10851 in this region is also assessed in terms of a Z-score calculated from the read depth in 100 bp bins.[4] After calling CNVs, CNVs which overlap with each other by more than 1 bp were collapsed into CNV regions and the midpoint of each CNV was set as its locus.

Candidate SNP genotyping for association test

Using TaqMan SNP Genotyping Assay (Applied Biosystems, Inc.), we genotyped 611 individuals, who were included in the family-based association study at the non-synonymous SNP (rs4148254). All the experiments were conducted according to the manufacturer's protocol. The sequences of primers and reporters are as follows; forward primer 5'-GTTCTTCCCAAATATGAAAGGATAATGCA-3', reverse primer 5'-GCTGGACCCATGAACCAAATCA-3', reporter 1 (VIC) 5'-TGGACTTCTCTGGCAGCAG-3', and reporter 2 (FAM) 5'-TGGACTTCTCTAGCAGCAG-3'.

Candidate copy number loss genotyping for association test

To determine the copy number status located 5.6 kb upstream (Chr4: 115 727 257-115 733 452) of *UGT8* (NM_001128174) using PCR, three kinds of primers were designed; (A) 5'-GCTCATGGATTGGAAGAACT-3', (B) 5'-CATGATCCTCTGATCCTCAAG-3', and (C) 5'-ACTGGCCAAGGGCTACTG-3'. The primers (A) and (B) are supposed to generate a product of 699 bp (Chr4: 115 727 028-115 727 726), which is only shown as a band in gel electrophoresis in the absence of copy number loss. The primers (A) and (C) are to target the region of 7017 bp (Chr4: 115 727 028-115 734 044), and these primers only work in the presence of copy number loss. These three primers were mixed and applied together to genomic DNA samples for genotyping which consists of two steps; PCR and gel electrophoresis. The PCR condition of this experiment was as follows; 30 s at 95 ℃ for denaturation, 30 s at 55 ℃ for annealing, and 60 s at 72 ℃ for amplification. We genotyped 618 individuals, who were analyzed in the family-based association study, for candidate copy number loss region**.**

3D model and motif analysis

We created a ribbon representation of a 3D model of the UGT8 protein using the MODELLER program [5] with the crystal structure of an *Acrocephalus orientalis* homolog (PDB: 1IIR [6]) as a template after checking the 3D-Jury [7] score. The picture was created using Swiss PDB Viewer [8]. For motif analysis, the ELM server was used to check if the loop with the Pro226 residue might contain a sequence motif that could be functionally important.

REFERENCES

1 Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010;**26**(7):873-81.

2 Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Church GM, Lee C, Kingsmore SF, Seo JS. A highly annotated whole-genome sequence of a Korean individual. *Nature* 2009;**460**(7258):1011-5.

3 Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* 2010;**42**(5):400-5.

4 Ju YS, Hong D, Kim S, Park SS, Lee S, Park H, Kim JI, Seo JS. Reference-unbiased copy number variant analysis using CGH microarrays. *Nucleic Acids Res* 2010;**38**(20):e190.

5 Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol* 2003;**374**:461-91.

6 Mulichak AM, Losey HC, Walsh CT, Garavito RM. Structure of the UDP-glucosyltransferase GtfB that modifies the heptapeptide aglycone in the biosynthesis of vancomycin group antibiotics. *Structure* 2001;**9**(7):547-57.

7 Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003;**19**(8):1015-8.

8 Kaplan W, Littlejohn TG. Swiss-PDB Viewer (Deep View). *Brief Bioinform* 2001;**2**(2):195-7.

LEGENDS TO FIGURES

Supplementary figure 1 Summary of exome sequencing for identifying causal variants. (A) Haploview LD plot of nsSNP (rs4148254) identified by exome sequencing for *UGT8*. Triangle color shows LD structure using D'/LOD score. r^2 values are indicated in the figure. The dark red color reflects a strong pairwise LD (r^2=0.8-1.0). The single and double asterisks indicate the top SNP from FBAT and the non-synonymous SNP from exome sequencing, respectively. (B) *UGT8* (NM_001128174) is composed of 6 exons including untranslated regions (white) and protein-coding sequence (gray). The variant (P226L) identified by exome sequencing is located in exon 2. The conserved domain, GT1_Gtf_like conserved domain (cd03784), ranges from amino acid positions 33 to 432. (C) The ribbon representation of a 3D model of the UGT8 protein. A crystal structure of the *Acrocephalus orientalis* homolog (PDB: 1IIR) was used as a template. The side chain of the key residue Pro226 is highlighted with a black circle.

Supplementary figure 2 The expression profile of *UGT8*. (A) Expression sequence tag profile of *UGT8* (UniGene; http://www.ncbi.nlm.nih.gov/UniGene). (B) Digital northern results of *UGT8* from serial analysis of gene expression (Cancer Genome Anatomy Project; http://cgap.nci.nih.gov/).