Supporting Information

Tilot et al. 10.1073/pnas.1715492115

SI Materials and Methods

Participants. Three families with multiple cases of sound–color (auditory–visual) synesthesia across several generations were identified from the Cambridge Synaesthesia Research Group database. The families were previously included as part of a 2009 study by Asher et al. (1), which utilized microsatellite markers to identify potential linkage regions across 43 families of various sizes. Exclusion criteria included a positive history of drug use or any neurological, ophthalmological, or psychiatric disorder. The authors confirmed synesthesia using the Test of Genuineness as described in the original study. Auditory stimuli consisted of both spoken words (e.g., days of the week, months, names, nouns) and other sounds (e.g., musical instruments, environmental sounds) (2). Ethical approval was granted by the Human Biology Research Ethics Committee of the University of Cambridge (Ref: 2011.06). Where original DNA samples from these families were depleted over the course of the 2009 study, consenting individuals were resampled using Oragene saliva-collection kits (DNA Genotek).

Sequencing. DNA was extracted from blood, buccal swabs, or Oragene kits of 18 individuals from the three selected synesthesia families (1). Library preparation began with 3 μg of purified DNA. Enrichment for exonic regions was performed using the SOLiD-optimized Sure Select All Human Exon Kit (50 Mb; Agilent Technologies) followed by sequencing on 5500XL sequencers (Life Technologies). Quality-control parameters were checked throughout the laboratory workflow. Sequence reads were aligned to the human genome (hg19) using Lifescope v2.1 (Life Technologies).

Variant Identification and Filtration. Variants were called using the best practices pipeline from the Genome Analysis Toolkit (GATK, version 2.8) (3). A further 56 exomes from other in-house studies (a total of 74 exomes) were included to improve calling accuracy. After marking duplicates, indel realignment, and base quality score recalibration (BQSR), variants were called from the 74 exomes using HaplotypeCaller (3). After variant calling, variant quality-score recalibration (VQSR) was performed for single-nucleotide variants and indels (3). Variants were annotated with Variant Effect Predictor (version 73), and GEMINI (version 0.18.3) was used for filtration based on segregation patterns and minor allele frequency (4, 5). Eigen scores were added using ANNOVAR version 2016-05-11 (wannovar.usc.edu/) (6, 7).

Sanger Validation. Rare variants that segregated with the synesthesia phenotype in each family were validated using Sanger sequencing. For the small subset of variants where DNA quality or quantity was insufficient for reliable PCR and Sanger sequencing, close inspection of the sequencing reads using Integrative Genomics Viewer (software.broadinstitute.org/software/igv/, version 2.3) was substituted. Representative images of these variants are included as Figs. S1 and S2.

Statistics and Data Visualization. Gene ontology analyses were performed with the Bioconductor package gProfileR (https://biit.cs.ut.ee/gprofiler/), using ontologies from Ensembl release 84, with a minimum set size of 10 genes and excluding inferred electronic annotations, and were single-tailed (we did not look for underrepresentation) (8). Neural gene-expression data were downloaded from GTEx, the Allen Human Brain Atlas, and the BrainSpan database, while mouse RNAseq data were accessed from Zhang et al. (9–11). All analyses and graphs were created in R (https://www.r-project.org/, version 3.1.1) using the additional packages ggplot2 (version 2.1.0), reshape2 (version 1.4.1), and ggthemes (version 3.0.3), with final figure production done in Inkscape (https://inkscape.org/, version 0.91). Network visualization was done in Cytoscape (www.cytoscape.org/, version 3.4.0), with functional interaction data from the Reactome FI plugin (apps.cytoscape.org/apps/reactomefiplugin, version 5.0.0).

Data Availability. The datasets generated during the current study are available upon request from The Language Archive (TLA: https://corpus1.mpi.nl/ds/asv/?0), a public data archive hosted by the Max Planck Institute for Psycholinguistics. The data are stored under the node IDs MPI1758324# and MPI1815362# and are accessible at https://hdl.handle.net/1839/00-0000-0000-001A-8756-4@view. All TLA content can be viewed from the Data Archiving and Networked Services Database, which is a Dutch national organization providing sustained access to digital research data. Allen Brain Institute data used in Fig. 2 *A* and *B* are available from BrainSpan (www.brainspan.org/static/download.html) and the Human Brain Atlas (human.brain-map.org/). Data used in Table S1 are available from GTEx (www.gtexportal.org/) and the Human Protein Atlas (v18.proteinatlas.org/). Data from the Barres laboratory used in Fig. 2*C* are available from https://web.stanford.edu/group/barres_lab/brain_rnaseq.html.

1. Asher JE, et al. (2009) A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12. *Am J Hum Genet* 84:279–285.
2. Asher JE, Aitken MR, Farooqi N, Kurmani S, Baron-Cohen S (2006) Diagnosing and phenotyping visual synaesthesia: A preliminary evaluation of the revised test of genuineness (TOG-R). *Cortex* 42:137–146.
3. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.1–11.10.33.
4. McLaren W, et al. (2016) The ensembl variant effect predictor. *Genome Biol* 17:122.
5. Paila U, Chapman BA, Kirchner R, Quinlan AR (2013) GEMINI: Integrative exploration of genetic variation and genome annotations. *PLOS Comput Biol* 9:e1003153.
6. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 92:841–853.
7. Yang H, Wang K (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* 10:1556–1566.
8. Reimand J, et al. (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* 44:W83–W89.
9. Battle A, Brown CD, Engelhardt BE, Montgomery SB; GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration & Visualization—EBI; Genome Browser Data Integration & Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017) Genetic effects on gene expression across human tissues. *Nature* 550:204–213, and erratum (2018) 553:530.
10. Miller JA, et al. (2014) Transcriptional landscape of the prenatal human brain. *Nature* 508:199–206.
11. Zhang Y, et al. (2014) An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* 34:11929–11947.

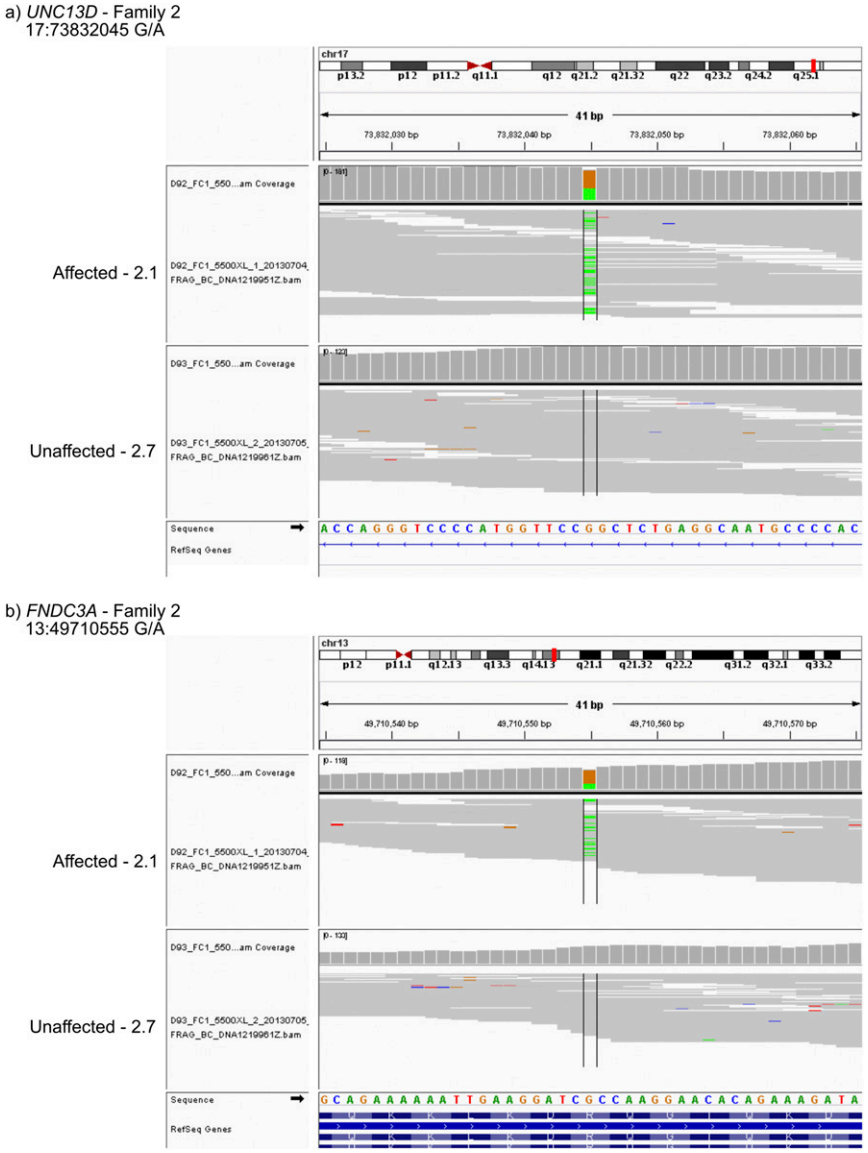

Fig. S1. Integrated Genomics Viewer images of variants in family 2 that could not be Sanger sequenced. (*A*) *UNC13D* 17:73832045 G/A. (*B*) *FNDC3A* 13:4971055 G/A.

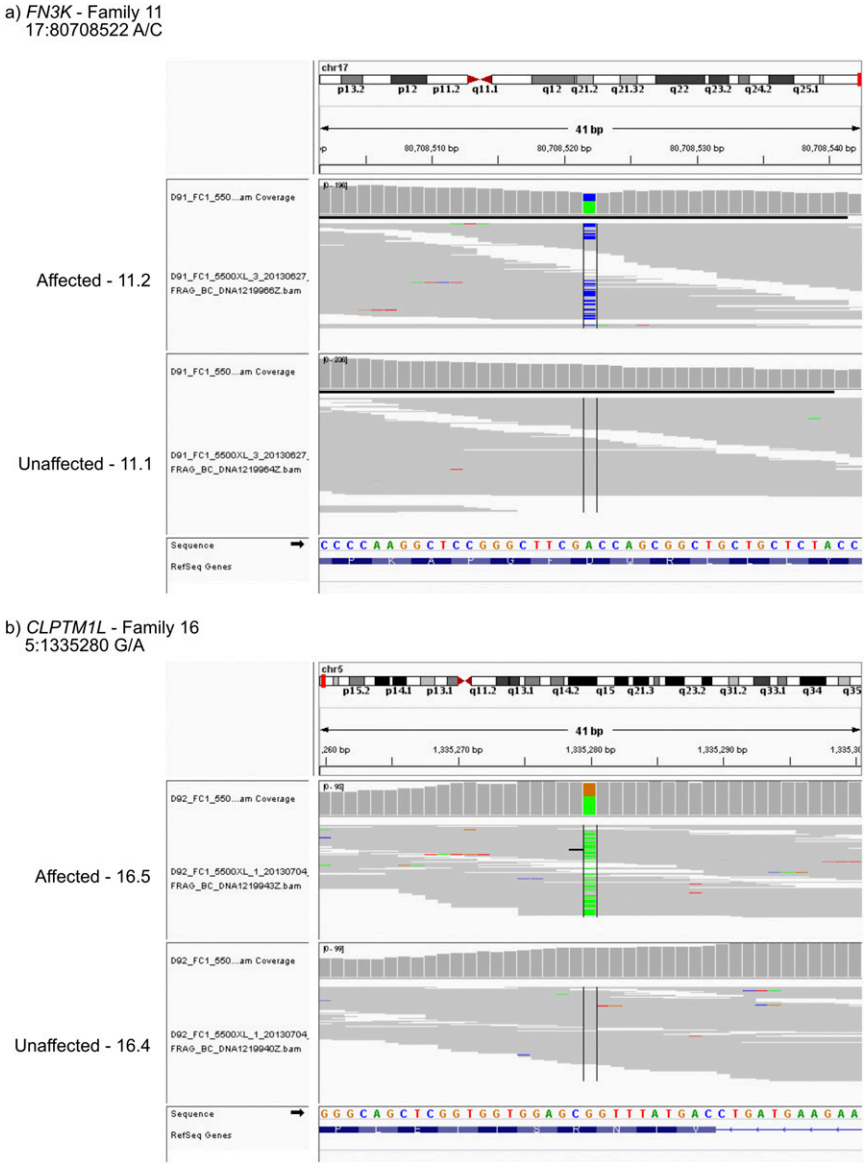

Fig. S2. Integrated Genomics Viewer images of variants that could not be Sanger sequenced. (*A*) Family 11, *FN3K* 17:80708522 A/C. (*B*) Family 16, *CLPTM1L*, 5:1335280.

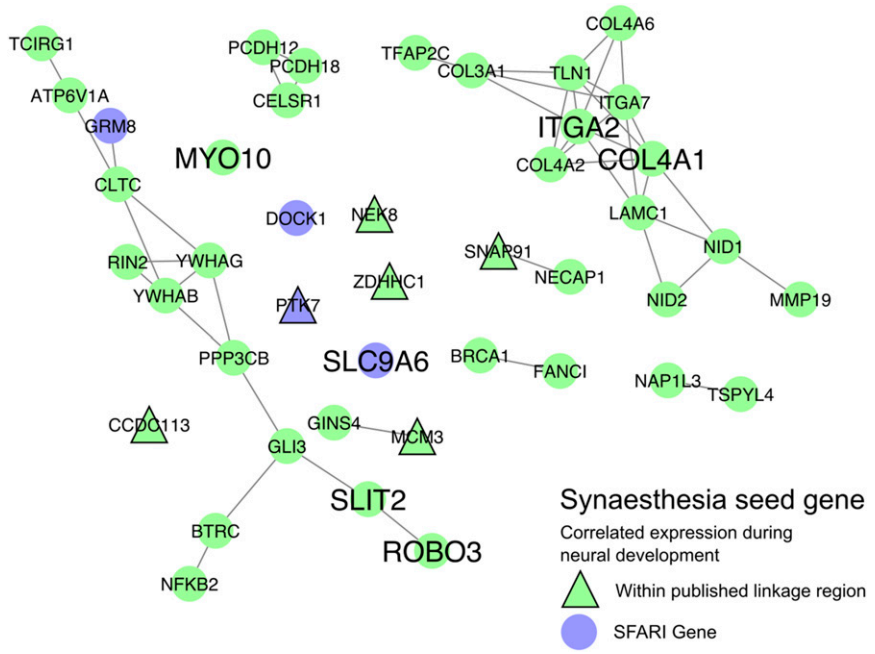

Fig. S3. Protein–protein interactions (edges) among genes (nodes) with neural expression correlated (small text) with one of the six synesthesia candidates (large text). Triangle-shaped nodes represent genes within previously published linkage regions; purple shading indicates genes in AutDB.

Table S1. Cortical RNA and protein levels for the top six candidates

	GTEx	Human Protein Atlas, antibody staining			
Gene	RNA, RPKM*	Endothelial cells	Glia	Neurons	Neuropil
COL4A1	2.33 ± 0.21	Medium	ND	Low	ND
ITGA2	0.87 ± 0.08	Low–medium	ND	ND–low	ND–medium
MYO10	13.54 ± 0.48	ND–high	ND–low	ND–high	Medium
ROBO3	6.21 ± 0.23	NA	NA	NA	NA
SLC9A6	17.74 ± 0.51	ND–low	ND–low	High	Medium
SLIT2	4.02 ± 0.14	ND–medium	ND–high	Medium–high	ND–medium

NA, protein levels not available in the Human Protein Atlas; ND, not detected.
*Values represent the mean ±1 SEM.