Imputation of Behavioral Candidate Gene Repeat Variants in 486,551 Publicly-Available UK Biobank Individuals.

Richard Border[1,2,3], Andrew Smolen[1], Robin P. Corley[1], Michael C. Stallings[1,2], Sandra A. Brown[4,5], Rand D. Conger[6], Jaime Derringer[7], M. Brent Donnellan[8], Brett C. Haberstick[1], John K. Hewitt[1,2], Christian Hopfer[1,9], Ken Krauter[1,10], Matthew B. McQueen[1,11], Tamara L. Wall[4], Matthew C. Keller[1,2], Luke M. Evans[1,12*]

[1] Institute for Behavioral Genetics, University of Colorado Boulder
[2] Department of Psychology and Neuroscience, University of Colorado Boulder
[3] Department of Applied Mathematics, University of Colorado Boulder
[4] Department of Psychiatry, University of California San Diego
[5] Department of Psychology, University of California San Diego
[6] Department of Human Ecology, University of California Davis
[7] Department of Psychology, University of Illinois at Urbana-Champaign
[8] Department of Psychology, Michigan State University
[9] Department of Psychiatry, University of Colorado Anschutz Medical Campus
[10] Department of Molecular and Cellular Biology, University of Colorado Boulder
[11] Department of Integrative Physiology, University of Colorado Boulder
[12] Department of Ecology and Evolutionary Biology, University of Colorado Boulder
* To whom correspondence should be addressed: luke.m.evans@colorado.edu

SUPPLEMENTARY MATERIAL

PubMed search results, Supplemental Tables S1-S7, Supplemental Figures S1-S3

PubMed 5HTTLPR Meta-analysis Search

On 5 June 2018, we identified 15 meta-analyses of the 5HTTLPR polymorphism in PubMed with the following search:

```
(("meta analysis"[Publication Type]) AND 5-HTTLPR) OR ("meta analysis"[Publication Type]
AND 5HTTLPR)) AND ("2015/01/01"[Date - Publication] : "2017/12/31"[Date - Publication]))
```

Due to the cursory nature of this search, this only provides a lower bound on the number of such studies during these years. The following papers were identified:

1. Bleys, D., Luyten, P., Soenens, B., & Claes, S. (2018). Gene-environment interactions between stress and 5-HTTLPR in depression: A meta-analytic update. *Journal of Affective Disorders*, *226*, 339–345.

2. Choi, H. D., & Shin, W. G. (2016). Meta-analysis of the association between a serotonin transporter 5-HTTLPR polymorphism and smoking cessation. *Psychiatric Genetics*, *26*(2), 87–91.

3. Clauss, J. A., Avery, S. N., & Blackford, J. U. (2015). The nature of individual differences in inhibited temperament and risk for psychiatric disease: A review and meta-analysis. *Progress in Neurobiology*, *127–128*, 23–45.

4. Gatt, J. M., Burton, K. L. O., Williams, L. M., & Schofield, P. R. (2015). Specific and common genes implicated across major mental disorders: a review of meta-analysis studies. *Journal of Psychiatric Research*, *60*, 1–13.

5. Li, H., Li, S., Wang, Q., Pan, L., Jiang, F., Yang, X., … Jia, C. (2015). Association of 5-HTTLPR polymorphism with smoking behaviors: A meta-analysis. *Physiology & Behavior*, *152*(Pt A), 32–40.

6. Mak, L., Streiner, D. L., & Steiner, M. (2015). Is serotonin transporter polymorphism (5-HTTLPR) allele status a predictor for obsessive-compulsive disorder? A meta-analysis. *Archives of Women's Mental Health*, *18*(3), 435–445.

7. Oo, K. Z., Aung, Y. K., Jenkins, M. A., & Win, A. K. (2016). Associations of 5HTTLPR polymorphism with major depressive disorder and alcohol dependence: A systematic review and meta-analysis. *The Australian and New Zealand Journal of Psychiatry*, *50*(9), 842–857.

8. Rozenblat, V., Ong, D., Fuller-Tyszkiewicz, M., Akkermann, K., Collier, D., Engels, R. C. M. E., … Krug, I. (2017). A systematic review and secondary data analysis of the interactions between the serotonin transporter 5-HTTLPR polymorphism and environmental and psychological factors in eating disorders. *Journal of Psychiatric Research*, *84*, 62–72.

9. Solmi, M., Gallicchio, D., Collantoni, E., Correll, C. U., Clementi, M., Pinato, C., … Favaro, A. (2016). Serotonin transporter gene polymorphism in eating disorders: Data from a new biobank and META-analysis of previous studies. *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry*, *17*(4), 244–257.

10. Suppli, N. P., Bukh, J. D., Moffitt, T. E., Caspi, A., Johansen, C., Albieri, V., … Dalton, S. O. (2015). 5-HTTLPR and use of antidepressants after colorectal cancer including a meta-analysis of 5-HTTLPR and depression after cancer. *Translational Psychiatry*, *5*, e631.

11. Taylor, S. (2016). Disorder-specific genetic factors in obsessive-compulsive disorder: A comprehensive meta-analysis. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *171B*(3), 325–332.

12. Tielbeek, J. J., Karlsson Linnér, R., Beers, K., Posthuma, D., Popma, A., & Polderman, T. J. C. (2016). Meta-analysis of the serotonin transporter promoter variant (5-HTTLPR) in relation to adverse environment and antisocial behavior. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *171*(5), 748–760.

13. Villalba, K., Attonito, J., Mendy, A., Devieux, J. G., Gasana, J., & Dorak, T. M. (2015). A meta-analysis of the associations between the SLC6A4 promoter polymorphism (5HTTLPR) and the risk for alcohol dependence. *Psychiatric Genetics*, *25*(2), 47–58.

14. Yamazaki, K., Yoshino, Y., Mori, T., Okita, M., Yoshida, T., Mori, Y., … Ueno, S.-I. (2016). Association Study and Meta-Analysis of Polymorphisms, Methylation Profiles, and Peripheral mRNA Expression of the Serotonin Transporter Gene in Patients with Alzheimer's Disease. Dementia and Geriatric Cognitive Disorders, 41(5–6), 334–347.

15. Zhao, Q., Guo, Y., Yang, D., Yang, T., & Meng, X. (2016). Serotonin Transporter Gene 5-HTTLPR Polymorphism as a Protective Factor Against the Progression of Post-Stroke Depression. *Molecular Neurobiology*, *53*(3), 1699–1705.

Supplemental Table S1. *DRD4* VNTR reciprocal reference imputation. The VNTR was assigned hg19 chromosome 11 physical position 640100 based on the UCSC genome browser. Risk alleles are 7 or more repeats. GP=genotype probability.

Target	Reference	True Risk Variant Freq.	Genotypes Used	Imputed Risk Variant Freq.	Minimac3 INFO score	Empirical r^2	Match Rate: Genotype	Match Rate: Allelic	Match Rate: Minor Allele		Number of alleles	Imputed: 0	Imputed: 1	Imputed: 2
CADD	FTP	0.207	All imputed genotypes	0.209	0.913	0.781	0.926	0.963	0.915	True	0	615	38	0
											1	25	302	3
											2	1	9	38
			GP>=0.99	0.182	-	0.961	0.988	0.994	0.979	True	0	521	2	0
											1	5	230	1
											2	0	1	25
FTP	CADD	0.198	All imputed genotypes	0.193	0.973	0.963	0.988	0.994	0.973	True	0	###	2	0
											1	15	597	0
											2	0	6	75
			GP>=0.99	0.174	-	0.977	0.993	0.996	0.981	True	0	###	0	0
											1	8	498	0
											2	0	4	54

Supplemental Table S2. *MAOA* reciprocal reference imputation in female individuals. The VNTR was assigned hg19 chromosome X physical position 43514400 based on the UCSC genome browser. Risk alleles were cosidered as 2, 3, or 5 repeats; 3.5 or 4 repeat alleles were considered wild type. GP=genotype probability.

Target	Reference	True Risk Variant Freq.	Genotypes Used	Imputed Risk Variant Freq.	Minimac3 INFO score	Empirical r^2	Match Rate: Genotype	Match Rate: Allelic	Match Rate: Minor Allele		Number of alleles	Imputed: 0	Imputed: 1	Imputed: 2
CADD	FTP	0.392	All imputed genotypes	0.388	0.965	0.574	0.802	0.890	0.855	True	0	94	17	3
											1	10	88	6
											2	3	15	37
			GP>=0.99	0.396	-	0.642	0.838	0.910	0.914	True	0	83	16	3
											1	6	81	5
											2	1	7	33
FTP	CADD	0.352	All imputed genotypes	0.339	0.946	0.819	0.921	0.959	0.923	True	0	419	22	0
											1	36	425	7
											2	3	14	115
			GP>=0.99	0.340	-	0.959	0.981	0.991	0.979	True	0	299	1	0
											1	6	310	2
											2	0	4	79

Supplemental Table S3. *MAOA* reciprocal reference imputation in male individuals. The VNTR was assigned hg19 chromosome X physical position 43514400 based on the UCSC genome browser. Risk alleles were cosidered as 2, 3, or 5 repeats; 3.5 or 4 repeat alleles were considered wild type. GP=genotype probability.

Target	Reference	True Risk Variant Freq.	Genotypes Used	Imputed Risk Variant Freq.	Minimac3 INFO score	Empirical r^2	Match Rate					Imputed	
							Genotype	Allelic	Minor Allele		Number of alleles	0	1
CADD	FTP	0.177	All imputed	0.160	0.973	0.567	0.888	-	0.791	True	0	343	22
											1	41	159
			GP>=0.99	0.159	-	0.665	0.919	-	0.854	True	0	331	18
											1	24	147
FTP	CADD	0.383	All imputed	0.348	0.934	0.791	0.947	-	0.884	True	0	544	8
											1	39	304
			GP>=0.99	0.353	-	0.954	0.989	-	0.970	True	0	411	1
											1	6	228

Supplemental Table S4. *SLC6A3* VNTR reciprocal reference imputation. The VNTR was assigned hg19 chromosome 5 physical position 1393863 based on the UCSC genome browser. Risk alleles are 10 or more repeats. GP=genotype probability.

Target	Reference	True Risk Variant Freq.	Genotypes Used	Imputed Risk Variant Freq.	Minimac3 INFO score	Empirical r^2	Match Rate					Imputed		
							Genotype	Allelic	Minor Allele		Number of alleles	0	1	2
CADD	FTP	0.762	All imputed genotypes	0.784	0.945	0.525	0.813	0.903	0.751	True	0	40	22	5
											1	22	250	92
											2	2	53	564
			GP>=0.99	0.800	-	0.728	0.898	0.949	0.857	True	0	34	11	0
											1	12	227	42
											2	0	25	535
FTP	CADD	0.757	All imputed genotypes	0.755	0.960	0.930	0.974	0.987	0.976	True	0	109	9	0
											1	6	708	14
											2	0	23	1113
			GP>=0.99	0.767	-	0.990	0.996	0.998	0.992	True	0	86	2	0
											1	0	596	4
											2	0	0	967

Supplemental Table S5. *SLC6A4* 5HTTLPR reciprocal reference imputation. The VNTR was assigned hg19 chromosome 17 physical position 28564497 based on the UCSC genome browser. Short alleles are 14 or fewer repeats, while 16 or more are considered long. GP=genotype probability.

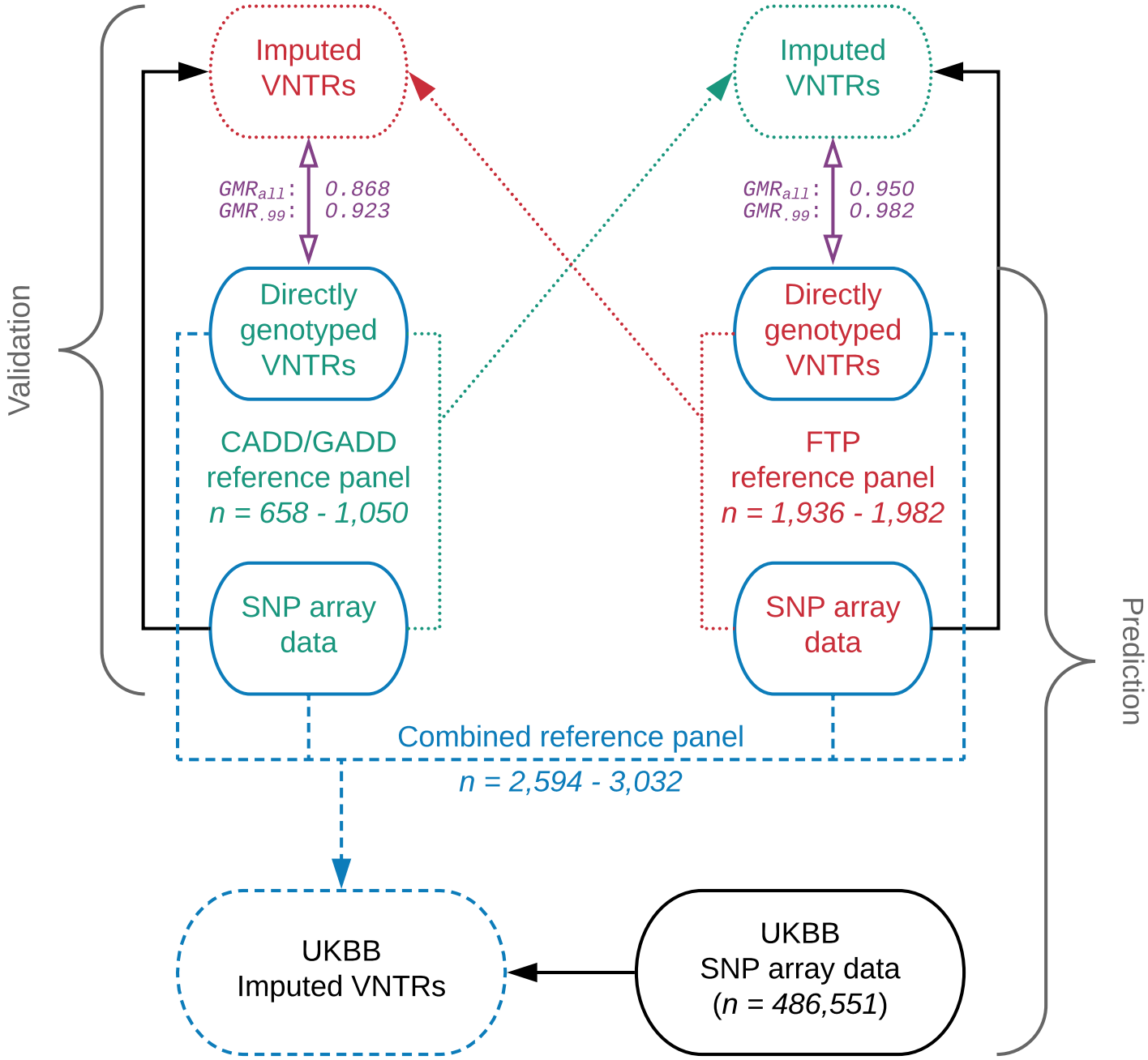
Target	Reference	True Long Allele Freq.	Genotypes Used	Imputed Long Allele Freq.	Minimac3 INFO score	Empirical r^2	Match Rate		Minor Allele		Number of alleles	Imputed		
							Genotype	Allelic				0	1	2
CADD	FTP	0.527	All imputed genotypes	0.527	0.926	0.692	0.842	0.917	0.908	True	0	214	51	6
											1	26	398	29
											2	3	51	274
			GP>=0.99	0.549	-	0.873	0.936	0.966	0.951	True	0	154	20	2
											1	5	302	9
											2	0	10	217
FTP	CADD	0.587	All imputed genotypes	0.591	0.940	0.834	0.919	0.959	0.946	True	0	302	32	3
											1	35	861	50
											2	0	38	642
			GP>=0.99	0.603	-	0.932	0.966	0.983	0.969	True	0	197	12	0
											1	5	559	19
											2	0	6	433

Supplemental Table S6. *SLC6A4* rs25531 (A/G) reciprocal reference imputation. The SNP is located on hg19 chromosome 17 at 28564346. GP=genotype probability.

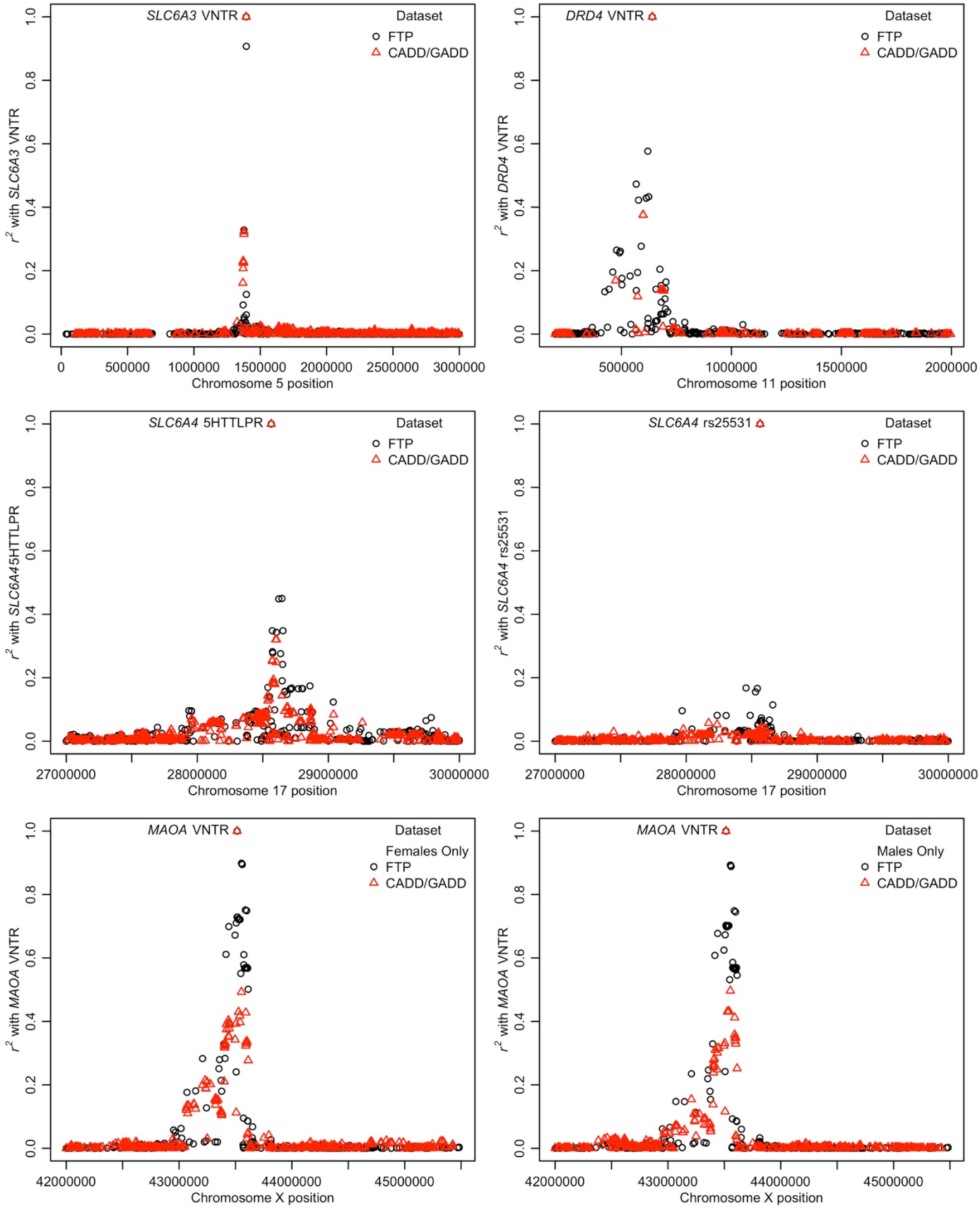
Target	Reference	True A Freq.	Genotypes Used	Imputed A Freq.	Minimac3 INFO score	Empirical r^2	Match Rate		Minor Allele		Number of alleles	Imputed		
							Genotype	Allelic				0	1	2
CADD	FTP	0.952	All imputed genotypes	0.937	0.952	0.474	0.935	0.967	0.794	True	0	2	0	0
											1	0	46	13
											2	0	30	567
			GP>=0.99	0.957	-	0.626	0.961	0.981	0.868	True	0	2	0	0
											1	0	42	7
											2	0	17	552
FTP	CADD	0.927	All imputed genotypes	0.930	0.974	0.658	0.951	0.976	0.813	True	0	3	4	0
											1	2	219	49
											2	0	40	1632
			GP>=0.99	0.939	-	0.737	0.967	0.984	0.835	True	0	2	2	0
											1	0	197	38
											2	0	20	1591

Supplemental Table S7. Imputation INFO score and imputed variant frequency from Minimac3 in the UK Biobank using the combined CADD+FTP reference panel. Imputation was performed in 4 randomly divided batches, with mean and standard deviation of INFO scores and frequency shown for each locus.

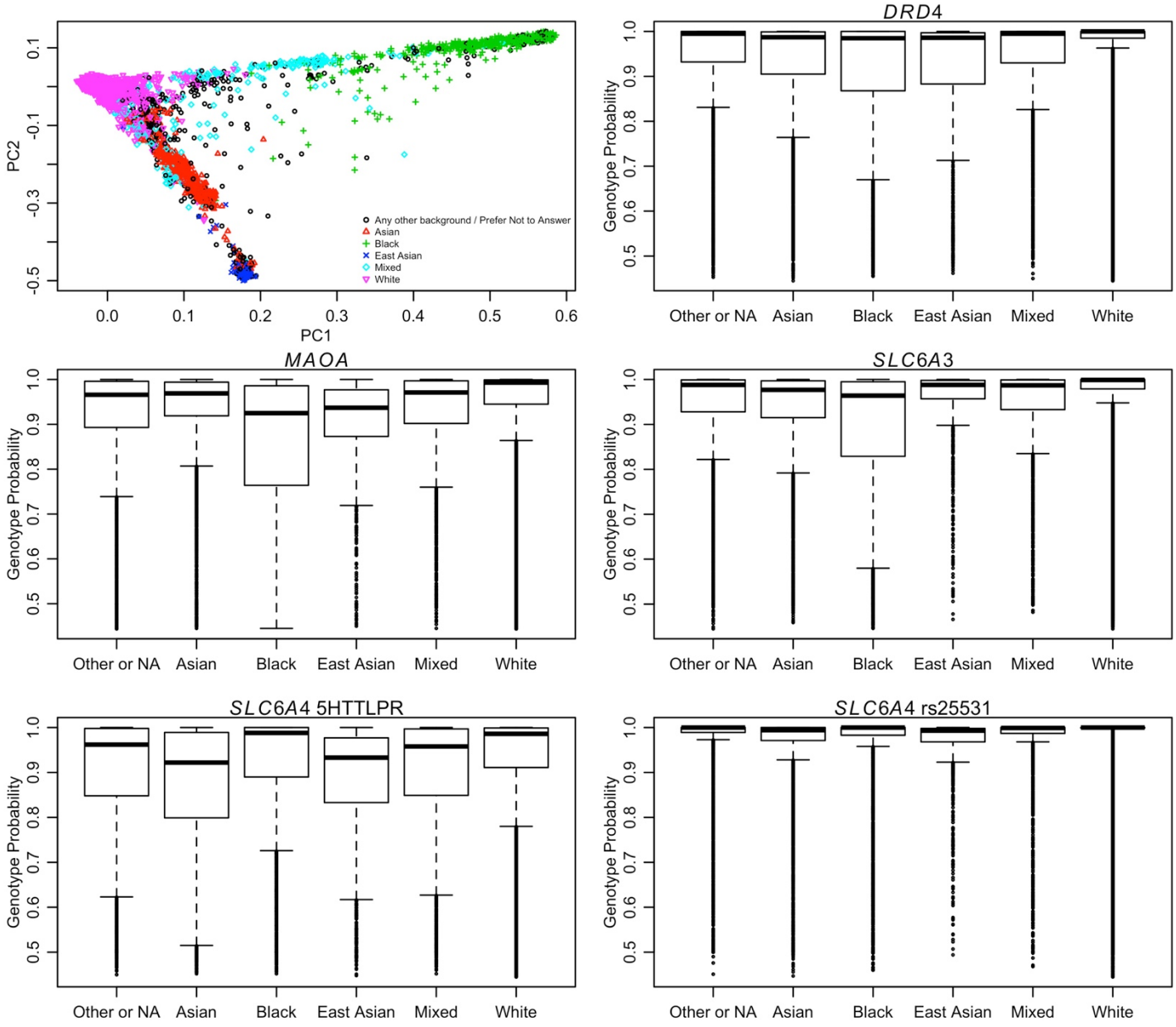
Locus	Batch	Chrom.	BP position	INFO r2	Mean	SD	Risk Var. Freq.	Mean	SD
SLC6A3 VNTR	a	5	1393863	0.9255	0.925	0.0004	0.253	0.253	0.0010
SLC6A3 VNTR	b	5	1393863	0.9253			0.252		
SLC6A3 VNTR	c	5	1393863	0.9249			0.254		
SLC6A3 VNTR	d	5	1393863	0.9258			0.254		
DRD4 VNTR	a	11	640100	0.9058	0.906	0.0010	0.212	0.211	0.0010
DRD4 VNTR	b	11	640100	0.9051			0.21		
DRD4 VNTR	c	11	640100	0.9054			0.211		
DRD4 VNTR	d	11	640100	0.9074			0.212		
MAOA VNTR	a females	X	43514400	0.9676	0.968	0.0003	0.639	0.639	0.0015
MAOA VNTR	a males	X	43514400	0.9682			0.640		
MAOA VNTR	b females	X	43514400	0.9681			0.641		
MAOA VNTR	b males	X	43514400	0.9677			0.639		
MAOA VNTR	c females	X	43514400	0.9678			0.640		
MAOA VNTR	c males	X	43514400	0.9684			0.638		
MAOA VNTR	d females	X	43514400	0.9679			0.640		
MAOA VNTR	d males	X	43514400	0.9684			0.636		
SLC6A4 rs25531	a	17	28564346	0.9094	0.907	0.0022	0.925	0.926	0.0006
SLC6A4 rs25531	b	17	28564346	0.9044			0.926		
SLC6A4 rs25531	c	17	28564346	0.9065			0.926		
SLC6A4 rs25531	d	17	28564346	0.9083			0.925		
SLC6A4 5HTTLPR	a	17	28564497	0.8850	0.883	0.0014	0.563	0.563	0.0008
SLC6A4 5HTTLPR	b	17	28564497	0.8829			0.563		
SLC6A4 5HTTLPR	c	17	28564497	0.8828			0.562		
SLC6A4 5HTTLPR	d	17	28564497	0.8816			0.564		



Supplemental Figure S1. Schematic of analyses performed to first validate our imputation strategy using reciprocally-imputed reference samples, then impute VNTRs into the UK Biobank using the combined reference panel. *n* is the sample size, GMR_{all} is the genotype match rate for all genotypes, and $GRM_{.99}$ is the genotype match rate for imputed genotypes with genotype probability of at least 0.99.



Supplemental Figure S2. Linkage disequilibrium, as measured by pairwise r^2, between each focal candidate gene variant and the surrounding array SNPs in the two reference samples.



Supplemental Figure S3. Comparison of the imputed genotype probability as a function of self-reported ethnic background (data-field 21000 in the UK Biobank). Top left PCA is coded by self-report ethnic background. Boxplots show medians and interquartile ranges, with whiskers extending to 1.5X the quartiles and more extreme observations shown as points. Genotype probability of imputed variants was significantly different among groups (one-way ANOVA, $F_{5,486545}>684$, $p<2x10^{-16}$). "Other or NA" indicates those of Any other background or who prefer not to answer.