Table of Contents

Ancestry Informative Markers

European ancestry was confirmed using TaqMan SNP assays for the 64 SNP marker subset of the 128 I_n4 ancestry informative marker (AIM) set defined by Seldin and colleagues[1]. Population admixture was initially evaluated using the program STRUCTURE v2.3.4 [2], run in the mixture mode assuming 3 major populations (Figure S-1).

The ancestry classification of individuals with predominantly European ancestry was refined by using the 64 AIM set to calculate a genetic distance for each individual to the same set of AIM genotypes from 938 individuals from the Human Genome Diversity Project (HGDP). The HGDP is a sample collection from 51 worldwide populations and the genotype data for the 64 AIM set was obtained from http://www.hagsc.org/hgdp/files.html. The genetic distance was calculated using the Ancestry Mapper package (CRAN: http://cran.r-project.org/), which represents distance by a vector (the Ancestry Mapper id or AMid) of normalized Euclidean distances for a SECASP individual to the 51 HGDP reference populations [3]. The SECASP AMids were clustered with the HGDP AMids using a Partitioning Around Medoids algorithm (pam function from the R package ‘cluster’) with the AMid as the metric and a cluster number of 50. By inspection of the distribution of SECASP participants within the HGDP European population groups (Figure S-2), participants in clusters 27 through 42 passed the study inclusion criteria for European ancestry.

Figure S-1. Admixture analysis of the SECASP participants, K =3.

The distance to each edge of the triangle gives the three components of the ancestry vector for an individual, estimated by STRUCTURE for three populations. Each SECASP individual is represented by a point and the color red indicates those included in this study.

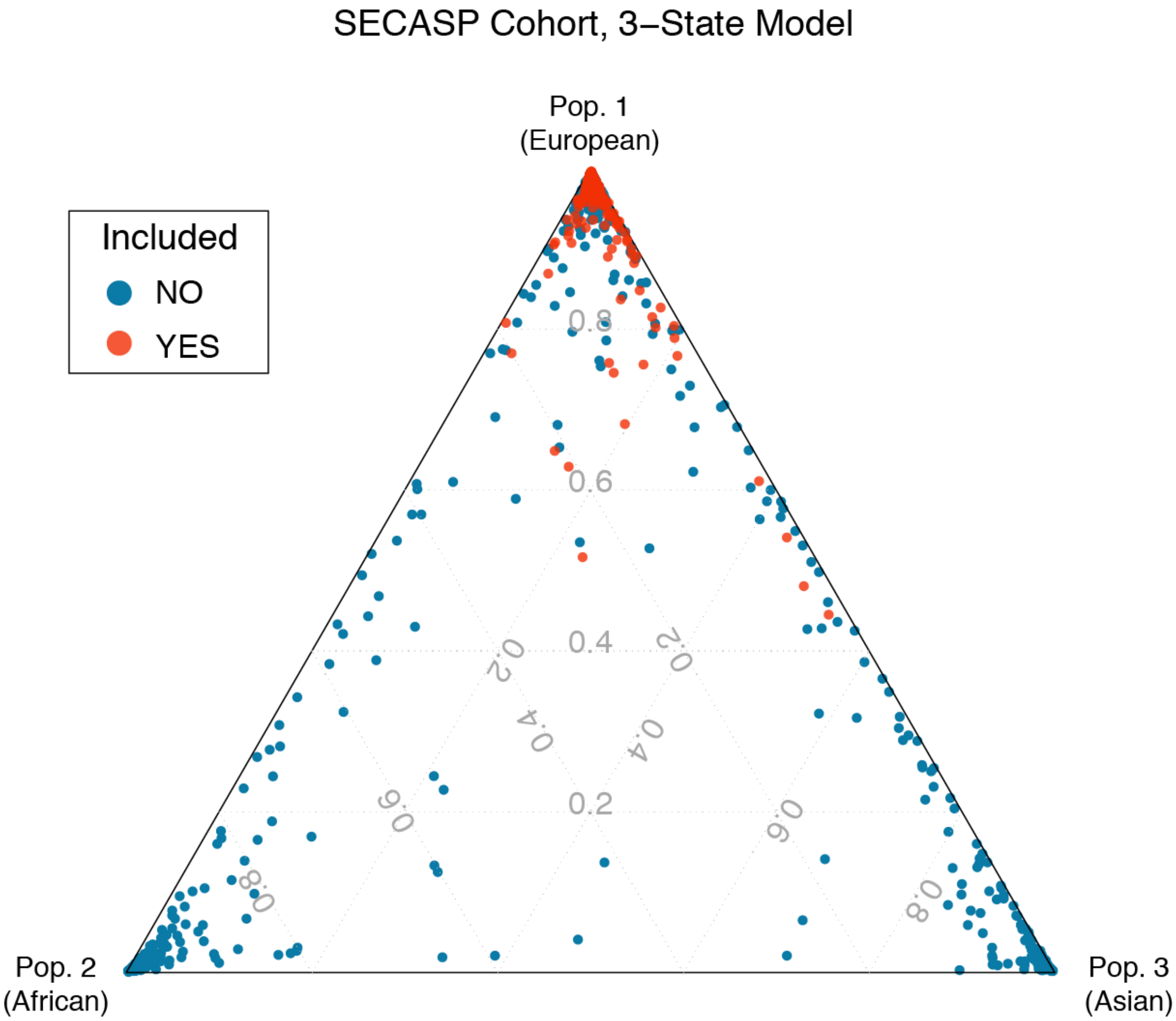

Figure S-2. Clustering of SECASP and HGDP individuals, K = 50.

All genotyped SECASP individuals (N=1019) are shown in 6 groups of self-reported race/ethnicity indicated by the vertical blue lines (NHB = Non-Hispanic black, HB = Hispanic blacks, NHW = Non-Hispanic whites, NHA = Non-Hispanic Asians, HW = Hispanic whites and OTHER = Other). The HGDP populations are arranged along the x-axis by their continental groupings (Africa, North Africa/Middle East, Europe, Central South Asia, Eastern Asia, America and Oceania). The matrix indicates the number of individuals in each cluster from these populations. Individuals in clusters 27 to 42 were assigned to the European ancestry group and the cluster assignments of SECASP individuals included in this study are highlighted in blue.

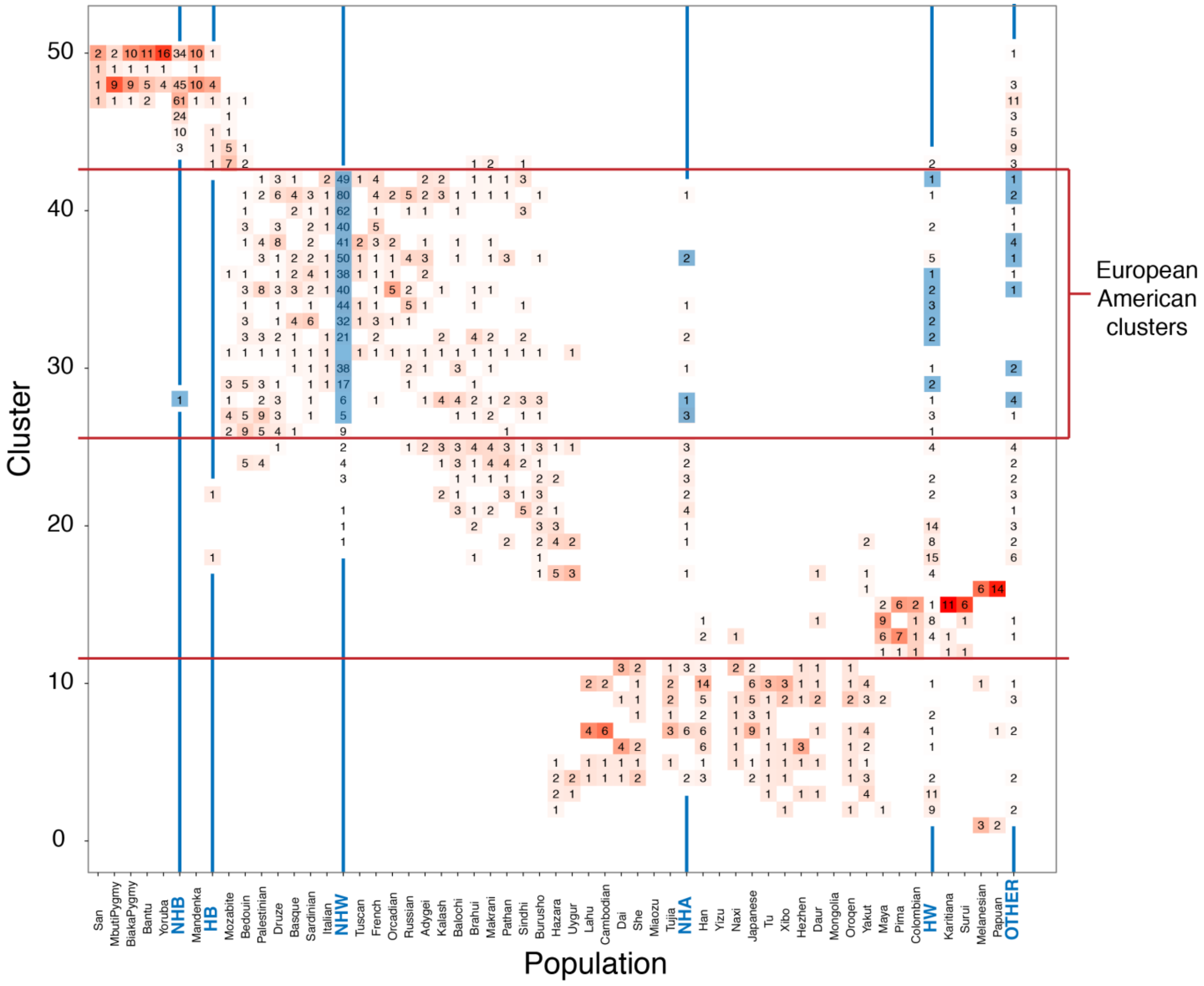

Table S-1. *CYP2A6* allele count and frequency data in the 296 novice smokers who had progressed to 100 cigarettes lifetime by Year 6.

Allele	Count	Frequency
Normal	373	0.63
*1A(51A)	128	0.22
*9	45	0.08
*12	27	0.045
*2	16	0.027
*4	2	<0.01
*1X2	1	<0.01
Total	296	1.00

***CYP2A6* Diplotype Predicted Rate (CDPR)**

CYP2A6 Diplotype Predicted Rate (CDPR) was based on *CYP2A6* diplotypes using a metric developed by Bloom and colleagues [4,5]. In cohorts of European descent, this metric ranges from a low of 0.44 for null metabolism alleles to 0.90 for normal metabolism alleles. The metric assigned to each diplotype in our implementation of Bloom's procedure is shown in Table S-2. Although this metric was intended to be used as a continuous variable [4], it was highly skewed in the study cohort, skew = -3.2 (cf. Figure S-3 and Table S-2). This negative skew attenuates its statistical associations, so CDPR was partitioned into 3 levels (cf. Table S-2).

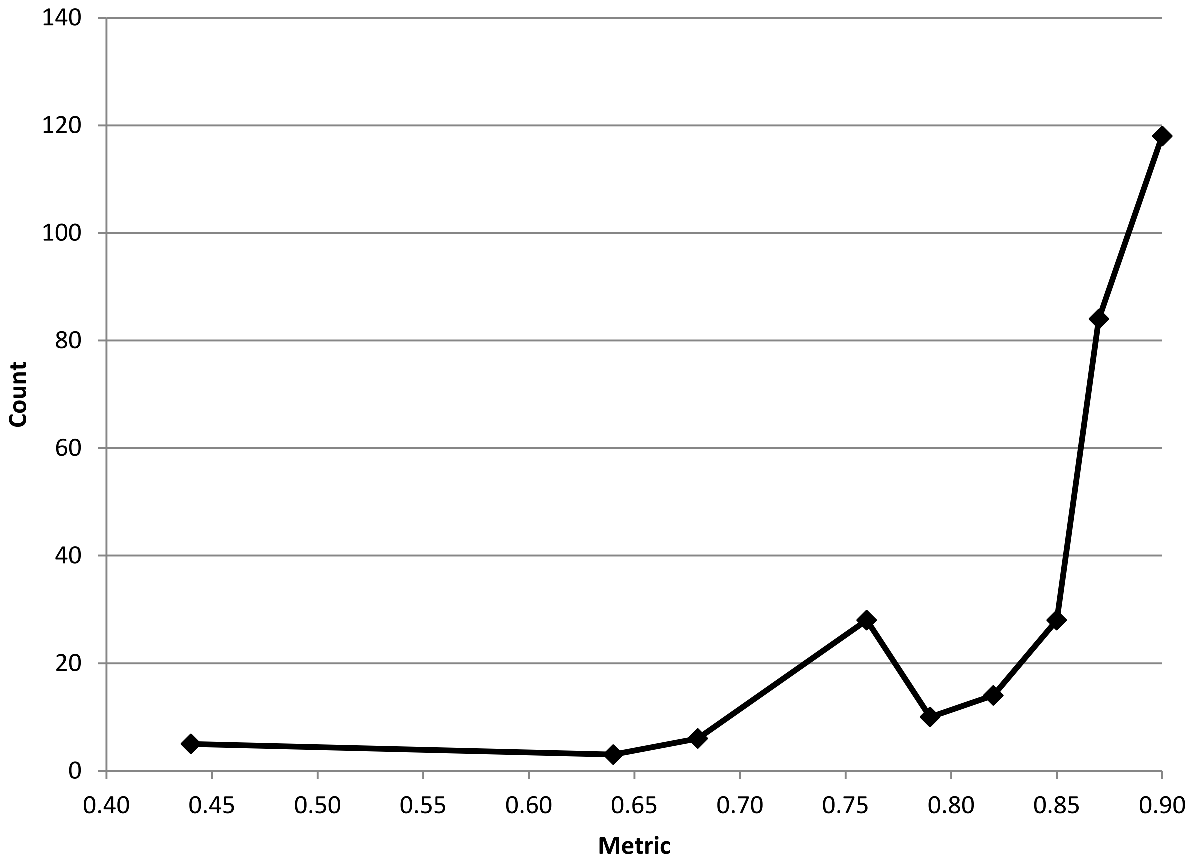

Figure S-3. Frequency distribution of metric values in the study cohort.

Table S-2. Diplotype metric value and frequency. "Normal" CDPR was defined as metric ≥ 0.87; "Intermediate" CDPR, metric ≥ 0.79 and < 0.87; and "Slow" CDPR, metric < 0.79. These cut-points are highlighted by showing Intermediate CDPR in gray.

Diplotype	Metric	Count	Frequency
Normal/Normal	0.90	118	0.40
Normal/*1A(51A)	0.87	83	0.28
*1A(51A)/*1X2	0.87	1	0.00
Normal/*9	0.85	28	0.09
*1A(51A)/*1A(51A)	0.82	14	0.05
*1A(51A)/*9	0.79	10	0.03
Normal/*12	0.76	19	0.06
Normal/*2	0.76	5	0.02
Normal/*4	0.76	2	0.01
*9/*9	0.76	2	0.01
*1A(51A)/*2	0.68	2	0.01
*1A(51A)/*12	0.68	4	0.01
*9/*12	0.64	2	0.01
*2/*9	0.64	1	0.00
*2/*2	0.44	4	0.01
*12/*12	0.44	1	0.00
Total		296	1.00

Table S-3. Growth-curve model predicting DAYS and NDSS, baseline through Year 6.

	DAYS						NDSS		
	Smoking Frequency (Poisson)			Smoking Discontinuation (Logistic)					
	Estimate	SE	P-Value	Estimate	SE	P-Value	Estimate	SE	P-Value
Intercept	1.767***	.052	4.8E-256	-.697**	.238	3.5E-03	1.260***	.111	9.0E-30
Time	.128***	.008	2.8E-52	-.131*	.055	1.7E-02	.133***	.026	4.5E-07
CDPR[a]									
Slow	-.139*	.059	1.8E-02	.885***	.263	7.6E-04	-.238*	.107	2.6E-02
Normal	-.152***	.042	2.6E-04	.320	.198	1.1E-01	-.158*	.078	4.3E-02
Time by CDPR[a]									
Time by Slow	.007	.013	6.1E-01	-.269**	.086	1.8E-03	.063	.040	1.1E-01
Time by Normal	.044***	.009	3.0E-06	-.115+	.062	6.4E-02	.072*	.029	1.4E-02
Controls									
Age at BL	-.028	.017	1.0E-01	.158*	.078	4.2E-02	-.160***	.047	7.5E-04
Male	.091***	.022	3.9E-05	-.010	.094	9.1E-01	.031	.057	5.9E-01
NHW	.013	.031	6.7E-01	.328+	.174	6.0E-02	.006	.097	9.5E-01
Random Effects									
Intercept	.173***	.010	1.7E-69				.133***	.018	3.9E-14
Slope							.020***	.003	1.1E-14
Intercept, Slope							-.006	.005	2.2E-01

Note: +=p<.10; *=p<.05; **=p<.005; ***p<.001. [a]=Intermediate CDPR is reference category. N=296 subjects with 2,274 observations over eight waves for DAYS, N=212 subjects with 1,374 observations over seven waves for NDSS. ZIP growth-curve model is used for DAYS and linear growth-curve model is used for NDSS. Residual variances (not shown) are freely estimated over time for linear growth-curve model.

Table S-4. Growth-curve model predicting DAYS in which Normal includes the *1A(51A) homozygote, Intermediate includes *1A(51A)/*9 and Normal/*9, and Slow is the same as in the primary CDPR classification.

	Smoking Frequency			Smoking Discontinuation		
	(Poisson)			(Logistic)		
	Estimate	SE	P-Value	Estimate	SE	P-Value
Intercept	2.083***	.056	7.1E-303	-1.141***	.274	3.1E-05
Time	.114***	.009	6.6E-34	-.031	.065	6.4E-01
CDPR[a]						
Slow	-.216**	.069	1.7E-03	1.334***	.294	7.4E-05
Normal	-.264***	.052	3.9E-07	.820**	.236	1.6E-03
Time by CDPR[a]						
Time by Slow	.019	.014	1.7E-01	-.369***	.093	5.5E-06
Time by Normal	.058***	.010	2.1E-08	-.224***	.071	5.1E-04
Controls						
Age at BL	-.034*	.016	3.4E-02	.160*	.078	3.9E-02
Male	.153***	.022	1.1E-11	-.013	.094	8.9E-01
NHW	-.230***	.032	4.2E-13	.326+	.174	6.1E-02
Random Effects						
Intercept	.172***	.009	3.6E-83			

Note: +=p<.10; *=p<.05; **=p<.005; ***p<.001. [a]=Intermediate CDPR is reference category.

N=296 subjects with 2,274 observations over eight waves.

Table S-5. Growth-curve model predicting DAYS using 2-category coding of CDPR, i.e., Intermediate and Slow were combined.

	Smoking Frequency (Poisson)			Smoking Discontinuation (Logistic)		
	Estimate	SE	P-Value	Estimate	SE	P-Value
Intercept	1.620***	.053	1.2E-201	-.323	.210	1.2E-01
Time	.133***	.007	1.0E-79	-.248***	.042	8.9E-59
CDPR[a]						
Normal	-.053	.051	3.0E-01	-.075	.157	6.3E-01
Time by CDPR[a]						
Time by Normal	.038***	.008	3.2E-06	.002	.051	9.7E-01
Controls						
Age at BL	.010	.029	7.2E-01	.152+	.078	5.1E-02
Male	-.003	.033	9.2E-01	-.001	.094	1.0E+00
NHW	.124***	.037	9.1E-04	.338+	.175	5.3E-02
Random Effects						
Intercept	.234***	.014	3.6E-09			

Note: +=p<.10; *=p<.05; **=p<.005; ***p<.001. [a]=Slow CDPR is reference category. N=296 subjects with 2,274 observations over eight waves for DAYS

Further discussion of participant selection criteria

Two smoking phenotype selection criteria were used: (1) participants had to have smoked at least a puff but fewer than 100 cigarettes lifetime at baseline, and (2) they had to have smoked at least 100 cigarettes lifetime by Year 6. The first criterion excluded both never-smokers and participants who by baseline had reached the level of lifetime nicotine exposure conventionally used to define "smokers" [6]. We here refer to participants included by the first criterion as "novice smokers." In our data, at baseline there were 458 novice smokers of European ancestry.

The second criterion selected from the 458 novice smokers at baseline the subset of 296 who had smoked at least 100 cigarettes lifetime by Year 6, i.e., in our terminology here, had progressed. This criterion was employed because we explicitly were interested in time-dependent within-participant changes in *CYP2A6* effects among participants who progressed. Thus, 162 novice smokers (458 – 296) were excluded by the second criterion.

While the combination of selection criteria identified the optimal cohort for the research question we did address, we here discuss several questions that might be raised relative to the second criterion. Specifically, did the second selection criterion bias the study cohort with respect to *CYP2A6* effects, and would the main effects observed in the study cohort have been observed had we included the full range of smoking progression/non-progression in the analyses?

1. *Was CYP2A6 associated with loss to follow-up at Year 6?*

Year 6 data were unavailable for 16 of the 458 baseline novice smokers. Having completed the Year 6 assessment was necessary for the application of the second selection criterion, so all 16 participants without Year 6 data were among the 162 participants excluded by criterion 2 and constitute 10% of the excluded novice smokers. A logistic regression analysis was based on the 162 excluded participants in which having Year 6 data was the dependent variable and sex, baseline age and self-reported race/ethnicity were covariates. There was no effect due to CDPR level. Thus, there is no evidence that loss to follow-up at Year 6 was biased with respect to *CYP2A6* alleles.

2. *Is CYP2A6 associated with smoking progression?*

Among the 442 novice baseline smokers with Year 6 data, 144 (33%) had not smoked 100 or more cigarettes lifetime by Year 6. To test *CYP2A6* effects on Year 6 lifetime cigarette exposure, 2 logistic regression models were tested in which having smoked 100+ cigarettes lifetime at Year 6 was the dependent variable and sex, baseline age and self-reported race/ethnicity were covariates. In one, Normal CDPR was the reference condition and in the other Intermediate CDPR was the reference condition. There were no CDPR effects in either model. Thus, we conclude that *CYP2A6* does not predict the progression of novice smokers to a lifetime exposure criterion of 100 cigarettes.

3. *Were the intercept effects observed for Intermediate CDPR on the smoking frequency portion of days smoked and for the NDSS contingent on excluding participants who did not progress?*

Cross-sectional baseline analyses were used to explore this question. In a ZIP model analysis of baseline days smoked with all 458 novice smokers and with sex, age and self-reported race/ethnicity as covariates, higher intercepts were observed for Intermediate CDPR relative to both Normal and Slow CDPR, p’s < .004, in the smoking frequency portion. For the 384 novice smokers with baseline NDSS, Intermediate CDPR was associated with higher scores than Normal CDPR, p = .02, but did not differ from Slow CDPR, p = .22. We conclude that the baseline risk effect of Intermediate CDPR relative to Normal CDPR is not dependent on limiting the analysis to novice smokers who progress. The Intermediate CDPR risk effect relative to Slow CDPR was observed for the smoking frequency portion of days smoked but not for the NDSS, suggesting this difference may be less robust.

4. *Was the intercept effect of Slow CDPR on smoking discontinuation contingent on excluding participants who did not progress?*

The association of Slow CDPR with smoking discontinuation was not observed when all novice smokers were considered. Inspection of baseline discontinuation probability by CDPR category suggests this attenuation of the protective effect of Slow CDPR is the result of greater discontinuation probability among novice smokers who did not progress but had either Normal or Intermediate CDPR. Discontinuation probability by CDPR for novice smokers who progressed and did not progress, respectively, was: Normal CDPR, 47%, 77%; Intermediate CDPR, 42%, 81%; and Slow CDPR, 64% and 65%.

5. *Was the Year 6 ordering of CDPR effects dependent on excluding participants who did not progress?*

In the smoking frequency portion of a ZIP model analysis of days smoked at Year 6 that included all baseline novice smokers, Normal CDPR was associated with more days smoked than either Intermediate or Slow CDPR, p's < .002, which is consistent with our findings for just the participants who progressed and with the reliable finding with adults. Year 6 NDSS genetic effects were not significant when all baseline novice smokers were considered, perhaps due to the universally low NDSS scores of those that did not progress.

Summary: We find no evidence that *CYP2A6* CDPR categories were associated with either loss to follow-up at Year 6 or with progression to 100 cigarettes lifetime by Year 6. We conclude that the baseline risk effect for Intermediate CDPR and the risk effect of Normal CDPR at Year 6 are observed when all novice smokers are considered. The protective baseline effect of Slow CDPR on discontinuation probability is not observed when all novice smokers are considered, but that most likely is due to increased baseline discontinuation in novice smokers with normal and intermediate metabolic rates who did not progress.

References

1. Kosoy R, Nassir R, Tian C, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 2009;30(1):69-78. doi: 10.1002/humu.20822.

2. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 2003;164(4):1567-1587. doi: PMC1462648.

3. Magalhaes TR, Casey JP, Conroy J, et al. HGDP and HapMap analysis by Ancestry Mapper reveals local and global population relationships. *PLoS One.* 2012;7(11):e49438. doi: 10.1371/journal.pone.0049438.

4. Bloom AJ, Harari O, Martinez M, et al. Use of a predictive model derived from in vivo endophenotype measurements to demonstrate associations with a complex locus, CYP2A6. *Hum. Mol. Genet.* 2012;21(13):3050-3062. doi: 10.1093/hmg/dds114.

5. Bloom J, Hinrichs AL, Wang JC, et al. The contribution of common CYP2A6 alleles to variation in nicotine metabolism among European-Americans. *Pharmacogenet. Genomics.* 2011;21(7):403-416. doi: 10.1097/FPC.0b013e328346e8c0.

6. Bierut LJ. Nicotine dependence and genetic variation in the nicotinic receptors. *Drug Alcohol Depend.* 2009;104 Suppl 1:S64-69. doi: 10.1016/j.drugalcdep.2009.06.003.