Nicotine & Tobacco Research, 2016, 626–631
doi:10.1093/ntr/ntv166
Brief report
Advance Access publication August 17, 2015

Brief report

Deep Sequencing of Three Loci Implicated in Large-Scale Genome-Wide Association Study Smoking Meta-Analyses

Shaunna L. Clark PhD[1], Joseph L. McClay PhD[1], Daniel E. Adkins PhD[1], Karolina A. Aberg PhD[1], Gaurav Kumar PhD[1], Sri Nerella MS[1], Linying Xie MS[1], Ann L. Collins PhD[2], James J. Crowley PhD[2], Corey R. Quakenbush MS[2], Christopher E. Hillard MS[2], Guimin Gao PhD[3], Andrey A. Shabalin PhD[1], Roseann E. Peterson PhD[4], William E. Copeland PhD[5], Judy L. Silberg PhD[4], Hermine Maes PhD[4], Patrick F. Sullivan MD, FRANZCP[2,6], Elizabeth J. Costello PhD[5], Edwin J. van den Oord PhD[1]

[1]Center for Biomarker Research and Precision Medicine, School of Pharmacy, Virginia Commonwealth University, Richmond, VA; [2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC; [3]Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA; [4]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA; [5]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Corresponding Author: Shaunna L. Clark, PhD, Center for Biomarker Research and Precision Medicine, School of Pharmacy, Virginia Commonwealth University, McGuire Hall, Room 216A, PO Box 980533, Richmond, VA 23298-0581, USA. Telephone: 804-628-3231; Fax: 804-628-3991; E-mail: slclark2@vcu.edu

Abstract

Introduction: Genome-wide association study meta-analyses have robustly implicated three loci that affect susceptibility for smoking: *CHRNA5\CHRNA3\CHRNB4*, *CHRNB3\CHRNA6* and *EGLN2\CYP2A6*. Functional follow-up studies of these loci are needed to provide insight into biological mechanisms. However, these efforts have been hampered by a lack of knowledge about the specific causal variant(s) involved. In this study, we prioritized variants in terms of the likelihood they account for the reported associations.

Methods: We employed targeted capture of the *CHRNA5\CHRNA3\CHRNB4*, *CHRNB3\CHRNA6*, and *EGLN2\CYP2A6* loci and flanking regions followed by next-generation deep sequencing (mean coverage 78×) to capture genomic variation in 363 individuals. We performed single locus tests to determine if any single variant accounts for the association, and examined if sets of (rare) variants that overlapped with biologically meaningful annotations account for the associations.

Results: In total, we investigated 963 variants, of which 71.1% were rare (minor allele frequency < 0.01), 6.02% were insertion/deletions, and 51.7% were catalogued in dbSNP141. The single variant results showed that no variant fully accounts for the association in any region. In the variant set results, *CHRNB4* accounts for most of the signal with significant sets consisting of directly damaging variants. *CHRNA6* explains most of the signal in the *CHRNB3\CHRNA6* locus with significant sets indicating a regulatory role for *CHRNA6*. Significant sets in *CYP2A6* involved directly damaging variants while the significant variant sets suggested a regulatory role for *EGLN2*.

Conclusions: We found that multiple variants implicating multiple processes explain the signal. Some variants can be prioritized for functional follow-up.

Introduction

Genome-wide association study meta-analyses, together comprising over 150 000 subjects,[1–4] have robustly implicated three loci that affect susceptibility for smoking behavior in subjects with European ancestry. These loci are: *CHRNA5\A3\B4* on chromosome 15, *CHRNB3\A6* on chromosome 8, and *EGLN2\CYP2A6* on chromosome 19. Functional follow-up studies of these loci are needed to provide insight into the biological mechanisms. However, these efforts have been hampered by a lack of knowledge about the specific causal variants involved. The goal of this article is to identify the variants likely to account for the previously detected associations in these loci. Given the robustness of these findings, we assume that a complete enumeration of all variants at each locus will likely contain the causal variants and our analyses can therefore be directed at prioritizing (sets of) variants based on statistical and bioinformatic evidence.

To achieve our goal, we employ targeted capture[5] of these loci in combination with deep, massively-parallel next-generation sequencing.[6] This approach is similar to exome sequencing,[7] but instead of capturing only exons, we capture the entire gene and their flanking regions. This method can identify all rare, low-frequency, and common single nucleotide polymorphisms (SNPs) plus small insertion/deletions. Our prioritization analyses examine whether a single common or low-frequency variant accounts for the association signal or if multiple (rare) variants are involved. In addition, by grouping variants in terms of biological function (eg, likely to affect regulatory function or protein coding) we can examine whether a single or multiple mechanisms are involved, which in turn allows us to begin generating functional hypotheses.

Methods

Sample and Measures

The data for this study come from the Virginia Twin Study on Adolescent and Behavioral Development (VTSABD[8]). A total of 363 independent individuals were included in the sequenced sample (ie, including only one twin from a twin pair). The sequenced sample was 38.8% male, 53.7% from a monozygotic twin pair, and 71.9% reported ever using a form of tobacco in their lifetime. The study was limited to subjects of European ancestry as insufficient numbers from other ancestry groups were available. Blood samples were collected when subjects were aged 25 to 34 from which DNA was extracted. All procedures were approved by ethical committees, and all subjects provided informed consent.

To be consistent with previous genome-wide association study meta-analyses that implicated the loci under consideration,[1–4] the phenotype used in this study is smoking quantity as measured by the number of cigarettes smoked per day. This is a prospective measure of the amount smoked per day over adolescence and early adulthood. For a further discussion of the phenotype, please see the Supplementary Material and Supplementary Table S1.

Sequencing

We used the solution-based hybridization targeted capture technology (SureSelectXT, Agilent) to target entire genes and ±5kb of their flanking regions. In this method, a library of synthetic oligonucleotides (baits) complementary to the sequence of interest is custom designed and manufactured.[5] These baits are then used to extract the desired genomic regions from fragmented genomic DNA samples. Library design and bait tiling were carried out using Agilent eArray. After removing overlap and collapsing neighboring genes into single loci, repetitive elements were removed as they can be difficult to align.

The libraries were paired-end sequenced (75bp + 35bp reads) on the SOLiD 5500xl (Life Technologies). The sequenced reads were aligned to the human genome (hg19/GRCh37) using Bioscope 1.3 (Life Technologies) that aligns in color-space and takes advantage of the increased ability of the SOLiD two-base encoding to identify sequencing errors. After alignment, quality control measures were implemented including dropping subjects with low mapped reads (<1 million) and fold enrichment (<365). Mean coverage across the targeted regions was 78×, with at least 10×/20× coverage for 95.4%/90.1% of the targeted regions, an average fold enrichment of 393.8, and 97.9% of baits covered. This level of coverage is very high for color space data, where two color call errors must occur by chance at the same position before a SNP is incorrectly called and therefore should result in fewer base calling errors relative to equivalent coverage on other sequencing platforms.[6]

Variant Calling and Annotation

The variants were called using GATK[9] using standard hard filtering parameters and variant quality score recalibration according to GATK Best Practices recommendations.[10,11] We defined rare variants as having a minor allele frequency < 0.01, low-frequency variants as 0.01 ≤ minor allele frequency < 0.05, and common variants as minor allele frequency ≥ 0.05. Singleton variants where the minor allele was found in one person were removed from the following analyses.

All variants passing quality control were annotated to examine if variants overlapped with bioinformatic features from the following databases: UCSC Genome Browser and GENCODE. Details are provided in Supplementary Table S2. To determine their novelty, identified variants were compared with dbSNP (v141[12]) and the 1000 Genomes Database[13] (1KG). Variants were also annotated for overlap with 15 chromatin states in brain tissue from the Anterior Caudate, Hippocampus, Mid-frontal Lobe, and Substantia Nigra regions, which are all known to be involved with nicotine addiction. See Supplementary Material for a description of the chromatin states and how they were generated by the Roadmap Epigenomics Project.[14] Annotations were used to prioritize variants and to form variant sets for the SNP set-based association tests.

Statistical Analyses

To prioritize (sets of) variants in terms of the likelihood they account for the reported associations, we used results from the tests described below.

Individual Variants

We performed single locus tests on all common and low-frequency variants passing quality control filters using a linear regression model of additive effects in PLINK.[15] Sex and 10 ancestry principal components (see Supplementary Material for a description of the principal component analysis) were included as covariates to control for sex differences and ancestry. Variants were then prioritized based on their effect size as measured by R^2, the square of the coefficient of multiple correlation, where $R^2 \geq 0.25$ is a large effect, $0.25 > R^2 \geq 0.09$ is a medium effect, $0.09 > R^2 \geq 0.01$ is a small effect, and $R^2 < 0.01$ is little to no effect.[16]

Sets of Variants

SNP set-based association tests were performed using SKAT[17] to prioritize sets of variants that may influence smoking. Specifically, we examined if biologically functional annotations (for a full list of annotations see Supplementary Table S2), such as variants that cause a damaging amino acid change as predicted by SIFT[18] and/or PolyPhen2,[19] for example, could account for the association signal with smoking. As in the single locus analysis, sex and 10 principal components were included as covariates. Since SKAT is meant to test association of sets of variants, and not estimate the effect size of a set,[20] the sets of variants were prioritized based on their *P* value. Through examining the types of variants sets that are significantly associated with smoking, we can begin to generate hypotheses about the potential mechanism through which these variants may influence smoking.

Results

Variant Calling

We identified 385 variants in the *CHRNA5\A3\B4* region (chr15:78,852,861-78,938,587) of which 279 (72.5%) were rare, 31 (8.05%) were insertion/deletions, 210 (54.5%) were catalogued in dbSNP141, 160 (41.6%) were in 1KG, and 115 (29.8%) were singletons. The *CHRNB3\A6* region (chr15:78,852,861-78,938,587) had 280 variants, of which 188 (67.1%) were rare, 12 (4.29%) were insertion/deletions, 157 (56.0%) were catalogued in dbSNP141, 130 (46.4%) were in 1KG, and 76 (27.1%) were singletons. There were 298 variants in the region of *EGLN2\CY2A6* (chr8:42,547,561-42,628,929) of which 218 (73.2%) were rare, 15 (5.03%) were insertion/deletions, 131 (43.9%) were catalogued in dbSNP141, 115 (38.6%) were in 1KG and 80 (26.9%) were singletons. For a complete list of variants and their annotations see Supplementary Table S3.

Individual Variants

The results of the single variant prioritization are shown in Table 1. In *CHRNB3\A6* and *EGLN2\CYP2A6*, there were only a few variants with $R^2 > 0.01$ while *CHRNA5\A3\B4* had many variants with $R^2 > 0.01$. However, none of the investigated loci had an individual variant with a large effect size. This suggests that, rather than a single variant with a large effect size, the previous genome-wide association study signals found in each region may have been caused by a variant we could not identify or by multiple variants with smaller effect sizes.

Sets of Variants

We investigated whether sets of all variants and only rare variants that overlapped with a biologically meaningful annotation were associated with smoking and could therefore potentially account for the previous association signals. The results with *P* value < .05 for each loci are shown in Table 2 and considered below.

CHRNA5\A3\B4

CHRNA5 had significant sets in evolutionarily conserved regions and repressed polycomb proteins, which have been shown to repress gene expression,[21] in three of the four brain regions. *CHRNB4* had nine significant sets that tended to be combinations of rare and non-rare sets. These sets included variants that can cause potentially damaging amino acid changes and heterochromatin, which is known to be involved in regulating gene expression.[22] These results suggest that significant sets of variants in *CHRNA5* and *CHRNB4* may have regulatory potential for smoking quantity. *CHRNA3* had no significant associations.

CHRNB3\A6

CHRNA6 had several significant sets of variants. The rare sets included missense variants, while the all-variant sets were related to repressed polycomb proteins. This suggests that significant variants sets in *CHRNA6* may have regulatory potential for smoking. *CHRNB3* had no significant set tests.

EGLN2\CYP2A6

Both *CYP2A6* and *EGLN2* had sets of variants overlapping with biologically meaningful annotations that were significantly associated with smoking. The significant sets in *CYP2A6* were variants that could potentially affect protein coding and chromatin states involved with the enhancement of transcription. The significant sets in *EGLN2* tended to be from rare variant sets with regulatory annotations like gene promoters, and chromatin states involving or flanking active transcription start sites. These results suggest that significant variants sets in *CYP2A6* may affect protein coding, while significant variant sets in *EGLN2* may have functional regulatory role.

Discussion

We prioritized variants in three loci known to be associated with smoking in order to identify variants that are likely to account for the association. The single variant results showed that no single common or low-frequency variant with a large effect size could fully explain the previous associations in any of the loci. There were, however, multiple variants with smaller effect sizes in each region that could account for some of the signal. This result is similar to what was seen in genome-wide association studies of smoking where variant(s) were found to be associated, but not to have a large effect size.[23] When considering sets of variants, we found evidence for overlap with potentially functional (affecting protein coding or regulatory) annotations. In *CHRNA5\A3\B4*, *CHRNB4* accounts for most of the signal with some significant sets consisting of damaging variants (ie, missense, POLYPHEN deleterious, and SIFT damaging) and variants that overlapped with heterochromatin. *CHRNA6* explains most of the signal in the *CHRNB3\A6* locus with significant sets indicating a potentially regulatory role for *CHRNA6*. Both *CYP2A6* and *EGLN2* could account for the association in that region, but through different processes. That is, the sets in *CYP2A6* tended to involve variants that may affect amino acid sequence in the encoded

Table 1. Common and Low-Frequency Variant Single Locus Results With $R^2 \geq 0.010$ by Loci

PSN(bp)	Gene	SNV number	RA	AA	AAF	EFF	*P*	R^2	Features
CHRNA5\A3\B4									
78892784	*CHRNA3*	rs62010327	G	A	0.373	–	.002	0.026	Intron, Shore
78897865	*CHRNA3*	rs75104798	C	CT	0.321	–	.004	0.023	Intron
78894971	*CHRNA3*	rs62010328	C	T	0.365	–	.005	0.022	Intron, Shore
78872211	*CHRNA5*		TCTTC	T	0.014	+	.006	0.021	Intron
78885988	*CHRNA5*	rs615470	T	C	0.623	–	.006	0.021	Exon
78909539	*CHRNA5*	rs3743073	G	T	0.623	–	.006	0.021	Intron, TFBScons
78881618	*CHRNA5*	rs17408276	T	C	0.375	–	.006	0.021	Intron
78887832	*CHRNA3*	rs660652	A	G	0.626	–		0.021	Exon
78869930	*CHRNA5*	rs495956	C	T	0.622	–	.007	0.020	Intron
78865694	*CHRNA5*	rs61012457	C	G	0.374	–	.010	0.018	Intron, TFBScons
78876505	*CHRNA5*	rs692780	C	G	0.626	–	.011	0.018	Intron
78856266		rs3829787	C	T	0.369	–	.014	0.017	Promoter(CHRNA5), Shore, TFBScons
78890321	*CHRNA3*	rs6495307	C	T	0.428	–	.021	0.015	Intron, TFBScons
78894896	*CHRNA3*	rs3743077	C	T	0.424	–	.021	0.015	Intron, Shore, TFBScons
78911780	*CHRNA3*	rs2067808	G	A	0.379	–	.022	0.015	Intron, Shore, DNase
78858491	*CHRNA5*	rs871058	G	A	0.355	–	.022	0.015	Intron, Shore, DNase
78884227	*CHRNA5*	rs514743	T	A	0.625	–	.030	0.013	Intron
78869579	*CHRNA5*	rs601079	T	A	0.572	–	.031	0.013	Intron
78871288	*CHRNA5*	rs386605197	T	C	0.572	–	.031	0.013	Intron
78865893	*CHRNA5*	rs6495306	G	A	0.573	–	.033	0.013	Intron
78907997	*CHRNA3*	rs11418931	A	AT	0.120	+	.034	0.012	Intron, TFBScons
78910267	*CHRNA3*	rs28669908	C	A	0.215	+	.050	0.011	Intron, Shore, DNase, TFBScons
78930510	*CHRNB4*	rs111358583	A	G	0.759	+	.057	0.010	Intron
CHRNB3\A6									
42598544		rs77112867	T	C	0.039	+	.018	0.016	
42563175	*CHRNB3*	rs4737066	A	G	0.966	+	.039	0.019	Intron
EGLN2\CYP2A6									
41316746		rs11668644	G	C	0.535	–	.022	0.019	Island, DNase, TFBScons, SuperDup
41349172		rs28742185	T	C	0.709	–	.041	0.012	SuperDup
41304074		rs117576995	G	A	0.011	+	.054	0.010	Promoter(EGLN2), Shore, DNase, TFBScons

AA = alternate allele; AAF = alternate allele frequency; PSN = position; RA = reference allele; SNV = single nucleotide variant. EFF is the direction of the effect of alternate allele where a "+" indicates smoking quantity is positively associated with alternate allele count and a "–" indicates smoking quantity is negatively associated with alternate allele count. "Feature" describes genomic attributes overlapping with the SNV's coordinates. "Exon" and "Intron" designate overlap with RefSeq genes; "DNase" indicates a genomic region hypersensitive to DNaseI; "Promoter" indicates the SNV is within 5kb of a transcription start site with the name of the gene it is promoting in parentheses; "CGI" denotes overlap with a CpG Island; "Shore" is ±2kb flanking a CGI; "SuperDup" designates overlap with a genomic super duplication; "TFBSCons" indicates SNV is within 100bp of a conserved transcription factor recognition sequence in mammals (TRANSFAC).

protein, while the significant rare variant sets suggested regulatory processes could underpin the association with *EGLN2*.

Several previous studies have sequenced the loci considered here, with multiple investigations having sequenced the *CHRNA5\A3\B4* and *CHRNB3\A6* loci as part of larger studies focusing on all cholinergic receptors with the goal of identifying causal variants for smoking.[24–28] These studies come to similar conclusions: that it is not a single variant acting alone that is causal, but rather sets of variants.[24,25,28] The sets they identified are rare variants that overlap with functional annotations such being missense or non-synonymous variants. Several of these previous studies focused only on exons,[24–26,29] and therefore would have missed 22 out of 27 (81.5%) of the single locus top findings and 23 out of 33 (69.6%) variant set top findings. To our knowledge, none of these previous investigations examined the role regulatory variant sets may play in smoking.

Our findings must be interpreted in the context of the potential limitations. A potential limitation is the statistical power for detecting sets of variants associated with smoking quantity with our modest sample size of 363. Using SKAT,[17] we conducted a small power study (see Supplementary Material for description and Supplementary Table S4 for results), which showed that only in a few extreme cases we would not have enough power to detect association with sets of variants. Additionally, given the strong prior associations of these regions with smoking, having enough power is less of an issue given that the goal is to prioritize variants that are already known to have an association. Another limitation is that our results suggest potential mechanisms through which the prioritized variants may affect smoking, rather than proving the mechanism. Possible next steps to do this include prioritization of these variants in an independent sample and examining the function of significant sets in targeted laboratory experiments. Potential methods include targeted genome

Table 2. Genomic Feature Set Results With *P* Value < .05

Gene	Feature	All *P* value	Rare *P* value	All vars.	Rare vars.
CHRNA3/CRNA5/CHRNB4					
CHRNB4	Heterochromatin—AC, HM, SN	4.71E-07	.017	4	1
CHRNB4	Heterochromatin—MFL	1.02E-04	.141	8	3
CHRNB4	Non Coding RNA	.009	.018	6	5
CHRNA5	Repressed PolyComb—SN	.012	.042	29	20
CHRNB4	Missense	.027	.034	6	3
CHRNB4	POLYPHEN—Deleterious	.036	.019	6	4
CHRNA5	Repressed PolyComb—MFL	.037	.228	50	34
CHRNA5	Repressed PolyComb—AC	.038	.495	35	25
CHRNB4	Conserved	.043	.015	6	3
CHRNA5	Conserved	.044	.052	3	1
CHRNB4	TFBS	.048	.082	2	2
CHRNB4	SIFT—Damaging	.054	.011	8	4
CHRNB4	Exon	.076	.023	13	7
CHRNB3/CHRNA6					
CHRNA6	Missense	1.75E-07	2.79E-07	5	5
CHRNA6	Conserved	5.57E-07	1.34E-06	10	10
CHRNA6	Exon	9.42E-05	2.69E-06	12	11
CHRNA6	Repressed PolyComb—HM	.002	.418	43	24
CHRNA6	Repressed PolyComb—AC	.030	.562	61	33
CHRNA6	Repressed PolyComb—SN	.067	.019	8	6
CYP2A6/EGLN2					
CYP2A6	Enhancer—HM	3.01E-04	.161	3	1
CYP2A6	Synonymous	.002	.037	10	5
CYP2A6	POLYPHEN—Deleterious	.003	.003	6	3
CYP2A6	SIFT—Damaging	.003	.005	8	4
CYP2A6	Exon	.016	.108	21	11
CYP2A6	Shore	.025	.153	30	19
EGLN2	Flank Active TSS—AC	.020	.006	14	7
EGLN2	Promoter	.033	.019	6	5
EGLN2	Enhancer—SN	.042	.390	36	17
EGLN2	Active TSS—HM	.073	.032	24	12
EGLN2	Active TSS—SN	.076	.039	25	13
EGLN2	Active TSS—MFL	.127	.011	34	19
EGLN2	Flank Active TSS—SN	.258	.016	17	8
EGLN2	Shore	.348	.003	80	42

"All *P* value" and "Rare *P* value" are the association *P* values from the test of whether the set of all variants or rare variants (minor allele frequency < 0.01) that overlap with the specified genomic feature within the given gene is associated with smoking. "All vars." and "Rare vars." is the number of variants included in the tested set. "Gene" indicates that the name of the gene the variant set falls within the boundary of as defined by RefSeq. "Feature" describes genomic attributes under consideration. "Conserved" indicates regions of high conservation across eutherian mammals; "Exon" designates overlap with RefSeq genes; "Missense" indicates the SNV is a missense mutation which results in an amino acid change; "Promoter" indicates the SNV is within 5kb of a transcription start site of the given gene; "Non Coding RNA" indicates a functional RNA molecule that is not translated into a protein; "Shore" is ± 2kb flanking a CpG Island; "Synonymous" indicates the variant is a coding single-nucleotide polymorphism that does not change the protein sequence; "TFBSCons" indicates SNV is within 100bp of a conserved transcription factor recognition sequence in mammals (TRANSFAC, [Matys et al., 2006]). "POLYPHEN—Deleterious" indicates that the variant is predicted to cause a deleterious amino acid substitution by PolyPhen2 [19]; "SIFT—Damaging" indicates that the variant is predicted to cause an amino acid substitution that is likely damaging to protein function by SIFT [18]. Chromatin states are indicated by the following format: chromatin state name—brain region of chromatin state set. Possible chromatin states are Active transcription start site (TSS), Flanking active TSS, Transcription at gene 5′ and 3′, Strong transcription, Weak transcription, Genic Enhancer, Enhancer, ZNF genes and repeats, Heterochromatin, Bivalent\Poised TSS, Flanking Bivalent TSS/Enhancer, Bivalent Enhancer, Repressed Polycomb, Weak Repressed Polycomb, and Quiescent. The brain regions examined for chromatin states were: AC—anterior caudate, HM—hippocampus, MFL—mid-frontal lobe, and SN—substantia nigra. Multiple brain regions listed in a single line indicate that the same set of variants formed the set for all listed brain region, hence the results are the same for these sets.

editing where DNA is changed using artificially engineered nucleases and the effect is observed,[30] or targeted chromatin immunoprecipitation assays of regulatory elements such as transcription factor binding sites and histone marks overlapping with the significant results.

In conclusion, we found that it is unlikely that a single common or low-frequency variant accounts for the entire association signal in any of the three smoking susceptibility loci considered. We identified specific genes within loci and specific sets of variants within those genes. This suggests it is likely that multiple variants and multiple processes are driving the association signal. We found interesting protein coding variant sets, however they do not account for all signals and it is likely that other variants also contribute via a regulatory role.

URLs

Agilent eArray: https://earray.chem.agilent.com/earray/
GENCODE: www.gencodegenes.org
UCSC Genome Browser: http://genome.ucsc.edu

Supplementary Material

Supplementary Material and Tables S1–S4 can be found online at http://www.ntr.oxfordjournals.org

Funding

This work was supported by the National Institutes of Health (R01 DA024413, R01 MH045268, R01 MH068521, R25 DA026119, K01 AA021266 to SLC, and K01 MH093731 to DEA).

Declaration of Interests

None declared.

References

1. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42(5):441–447. doi:ng.571 [pii] 10.1038/ng.571.
2. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet*. 2010;42(5):448–453. doi:ng.573 [pii] 10.1038/ng.573.
3. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet*. 2010;42(5):436–440. doi:ng.572 [pii] 10.1038/ng.572.
4. Saccone NL, Culverhouse RC, Schwantes-An TH, et al. Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet*. 2010;6(8). doi:10.1371/journal.pgen.1001053 [pii] e1001053.
5. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009;27(2):182–189. doi:nbt.1523 [pii]: 10.1038/nbt.1523.
6. McKernan KJ, Peckham HE, Costa GL, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*. 2009;19(9):1527–1541. doi:gr.091868.109 [pii]: 10.1101/gr.091868.109.
7. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009;461(7261):272–276. doi:nature08250 [pii]: 10.1038/nature08250.
8. Meyer JM, Silberg JL, Simonoff E, et al. The Virginia Twin-Family Study of Adolescent Behavioral Development: assessing sample biases in demographic correlates of psychopathology. *Psychol Med*. 1996;26(6):1119–1133.
9. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–1303. doi:gr.107524.110 [pii]: 10.1101/gr.107524.110.
10. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;11(1110):11.10.11–11.10.33. doi:10.1002/0471250953.bi1110s43.
11. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491–498. doi:ng.806 [pii]: 10.1038/ng.806.
12. Phillips C. Online resources for SNP analysis: a review and route map. *Mol Biotechnol*. 2007;35(1):65–97. doi:MB:35:1:65 [pii].
13. Clarke L, Zheng-Bradley X, Smith R, et al. The 1000 Genomes Project: data management and community access. *Nat Methods*. 2012;9(5):459–462. doi:nmeth.1974 [pii]: 10.1038/nmeth.1974.
14. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317–330. doi:nature14248 [pii]: 10.1038/nature14248.
15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575. doi:S0002-9297(07)61352-4 [pii]: 10.1086/519795.
16. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
17. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762–775. doi:kxs014 [pii]: 10.1093/biostatistics/kxs014.
18. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4(7):1073–1081. doi:nprot.2009.86 [pii]: 10.1038/nprot.2009.86.
19. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248–249. doi:nmeth0410-248 [pii]: 10.1038/nmeth0410-248.
20. Lee S. SKAT User Group Discussion Board. 2014. https://groups.google.com/forum/#!topic/skat_slee/9BklI9n-H1w. Accessed January 12, 2015.
21. Di Croce L, Helin K. Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol*. 2013;20(10):1147–1155. doi:nsmb.2669 [pii]: 10.1038/nsmb.2669.
22. Grewal SI, Jia S. Heterochromatin revisited. *Nat Rev Genet*. 2007;8(1):35–46. doi:nrg2008 [pii]: 10.1038/nrg2008.
23. Loukola A, Hallfors J, Korhonen T, et al. Genetics and smoking. *Curr Addict Rep*. 2014;1(1):75–82. doi:10.1007/s40429-013-0006-3.
24. Haller G, Druley T, Vallania FL, et al. Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. *Hum Mol Genet*. 2012;21(3):647–655. doi:ddr498 [pii]: 10.1093/hmg/ddr498.
25. Xie P, Kranzler HR, Krauthammer M, et al. Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biol Psychiatry*. 2011;70(6):528–536. doi:S0006-3223(11)00435-5 [pii]: 10.1016/j.biopsych.2011.04.017.
26. Wessel J, McDonald SM, Hinds DA, et al. Resequencing of nicotinic acetylcholine receptor genes and association of common and rare variants with the Fagerstrom test for nicotine dependence. *Neuropsychopharmacology*. 2010;35(12):2392–2402. doi:npp2010120 [pii]: 10.1038/npp.2010.120.
27. Weiss RB, Baker TB, Cannon DS, et al. A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. *PLoS Genet*. 2008;4(7):e1000125. doi:10.1371/journal.pgen.1000125.
28. Yang J, Wang S, Yang Z, et al. The contribution of rare and common variants in 30 genes to risk nicotine dependence [advance online publication December 2, 2014]. *Mol Psychiatry*. 2014. doi:mp2014156 [pii]: 10.1038/mp.2014.156.
29. Saccone NL, Wang JC, Breslau N, et al. The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res*. 2009;69(17):6848–6856. doi:0008-5472.CAN-09-0786 [pii]: 10.1158/0008-5472.CAN-09-0786.
30. de Souza N. Primer: genome editing with engineered nucleases. *Nat Methods*. 2012;9(1):27.