Published in final edited form as:

Pharmacol Biochem Behav. 2014 August ; 0: 34–44. doi:10.1016/j.pbb.2014.01.009.

Finding Genomic Function for Genetic Associations in Nicotine Addiction Research: The ENCODE Project's Role in Future Pharmacogenomic Analysis

David J. Vandenberg^{a,b,*} and Gabriel L. Schlomer^c

^aDepartment of Biobehavioral Health 219 Biobehavioral Health Building The Pennsylvania State University University Park, PA 16802 United States of America

^bPenn State Institute of the Neurosciences 101 Life Sciences Building University Park, PA 16802 United States of America

^cDepartment of Human Development and Family Studies The Pennsylvania State University 315 Health and Human Development – East University Park, PA 16802 United States of America
gls29@psu.edu

Abstract

Tobacco-related behaviors and the underlying addiction to nicotine are complex tangles of genetic and environmental factors. Efforts to understand the genetic component of these traits have identified sites in the genome (Single Nucleotide Polymorphisms, or SNPs) that might account for some part of the role of genetics in nicotine addiction. Encouragingly, some of these candidate SNPs remain significant in meta-analyses. However, genetic associations cannot be fully assessed, regardless of statistical significance, without an understanding of the functional consequences of the alleles present at these SNPs. The proper experimental test for allelic function can be very difficult to define, representing a roadblock in translating genetic results into treatment to prevent smoking and other nicotine-related behaviors. This roadblock can be navigated in part with a new web-based tool, the Encyclopedia of DNA Elements (ENCODE). ENCODE is a compilation of searchable data on several types of biochemical functions or “marks” across the genome. These data can be queried for the co-localization of a candidate SNP and a biochemical mark. The presence of a SNP within a marked region of DNA enables the generation of better-informed hypotheses to test possible functional roles of alleles at a candidate SNP. Two examples of such co-localizations are presented. One example reveals ENCODE's ability to relate a candidate SNP's function with a gene very far from the physical location of the SNP. The second example reveals a new potential function of the SNP, rs4105144, that has been genetically associated with the number of cigarettes smoked per day. Details for accessing the ENCODE data for this SNP are provided to serve as a tutorial. By serving as a bridge between genetic associations and

© 2014 Elsevier Inc. All rights reserved

*Corresponding Author djv4@psu.edu [+1 814 863 8430].

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

CONFLICTS OF INTEREST The authors have no conflicts to describe.

biochemical function, ENCODE has the power to propel progress in untangling the genetic aspects of nicotine addiction—a major public health concern.

Keywords

ENCODE; smoking; endophenotype; addiction; habenula; GWAS; CYP2A6; CHRNA5; CHRNA3; CHRNA4; rs4105144

1. Introduction

The role of genetics in tobacco-related traits, primarily smoking and nicotine addiction, has gained attention due to recent associations between these traits and SNPs in a cluster of 3 of the nicotinic acetylcholine receptor subunit genes, *CHRNA5*, *CHRNA3*, and *CHRNA4*, on chromosome 15q25 (abbreviated as *A5-A3-B4*) (Saccone *et al.*, 2007; Thorgeirsson *et al.*, 2008; Berrettini *et al.*, 2008). These findings are exciting for three reasons. First, they have been replicated in multiple studies (Liu *et al.*, 2010; Furberg *et al.*, 2010; Thorgeirsson *et al.*, 2010). Second, the findings are consistent with animal studies that demonstrate some of the brain regions in which these receptor subunit genes are expressed play a role in behavioral responses to nicotine (see Picciotto and Kenny, 2013 for review). Third, and perhaps most important, some of the SNPs are likely to alter the function of the gene in which they occur (Bierut *et al.*, 2008).

Assessing the potential for genetic variation to alter biological function is a major goal of genetic research on tobacco-related behaviors and nicotine addiction, which is caused by the psychoactive effects of nicotine. The functional consequences of some SNPs can be assigned when the two alleles of the SNP encode different amino acid sequences of the protein produced by the gene (known as non-synonymous coding SNPs) (*e.g.* Bierut *et al.*, 2008). For many of the SNPs, however, there is no change in protein sequence, and thus no clear hypothesis of function to test. There is a wide range of biochemical functions that might be proposed for a candidate SNP, and some of these functions, such as altering DNA methylation, histone modification, or accessibility of DNA to transcription factor binding, can have significant phenotypic effects by altering when, where, and how much a gene, and its protein, is expressed. Until recently, these types of functions have not been easy to predict, and so have not been tested for many candidate SNPs.

This review focuses on a new tool, the Encyclopedia of DNA Elements (ENCODE), that allows predictions to be made for the role a SNP might play in altering genomic function. Alterations in genomic function can provide a mechanism to explain how changes in DNA sequence are associated with behavioral phenotypes.

In order to demonstrate ENCODE's potential role in proposing function for candidate SNPs, genetic association studies of the *A5-A3-B4* cluster of the nicotinic receptor subunit genes, as well as the *CYP2A6* gene on chromosome 19, are discussed. Recent meta-analyses suggest robust, replicable findings for these SNPs. It seems clear that these regions of the genome impact tobacco-related behaviors in some way. In the case of the *A5-A3-B4* cluster, some of the excitement is due to data that suggest some of the SNPs in this region alter

function of one or more of the genes and that these changes are important in brain regions that regulate nicotine-related behaviors. While experiments directed at the *A5-A3-B4* cluster are making headway on understanding the function of SNPs in the cluster, there are other SNPs, such as the candidate SNP near the *CYP2A6* gene, whose functional consequences are largely unknown, despite robust associations with tobacco-related behaviors. This *CYP2A6*-related SNP will be discussed as a contrasting example to the SNPs in the *A5-A3-B4* cluster, and will demonstrate a way to overcome the technical challenges of assessing function of SNPs.

After providing details on recent genetic studies, the review briefly describes ENCODE, a web-accessible database that can reveal possible functional links between SNPs and their associated phenotypes. The first of two examples from ENCODE highlights its ability to detect a functional relationship between a SNP and a gene when the two are far apart on a chromosome. The second example is relevant to tobacco-related behaviors, and serves as a practical guide for using ENCODE. In this second case, a new hypothetical function is proposed for the SNP near *CYP2A6* that might help explain the cause of its association with a tobacco-related phenotype. As further genetic associations with tobacco-related behaviors are made, it is likely that ENCODE will be a source to provide a deeper knowledge of SNP functions, which may lead to better treatments for smoking—the largest single cause of premature death in the United States (CDC, 2004), and a major health problem world-wide (WHO, 2011; Giovino *et al.*, 2012).

2. Maturation of genetic analysis of tobacco-related behaviors

The recent genetic findings that some of the nicotinic acetylcholine receptor subunit genes are associated with tobacco-related phenotypes serve as good examples of significant changes in the field of tobacco-related genetics in the past two decades. One change has been an increased focus on the SNP-type of polymorphism, because of their frequency in the genome and the robustness with which they can be tested. There are many other types of polymorphisms involving changes as small as two bases (*e.g.* Simple Tandem Repeats) to as large as many thousands of bases (*e.g.* large deletions or insertions), but all of the studies described in this review are focused on SNPs.

Early genetic association studies generally focused on a single SNP in part because very few polymorphisms were known, and testing them required large (microgram) amounts of DNA from each study participant (Comings *et al.*, 1996; Lerman *et al.*, 1997). The field changed significantly when genotyping technology could detect many more polymorphisms per subject with a thousand-fold less DNA. As additional SNPs were identified, studies began incorporating several genes in a single study, although, each gene was still usually defined by a single SNP (Vandenbergh *et al.*, 2007). Researchers also began to appreciate that the genetics underlying tobacco-related traits is complex, which spurred the field toward multiple gene testing. Further leaps in genotyping technology allowed the more recent major advancement in which the entire genome's contribution in a Genome-Wide Association Study (GWAS) could be applied to smoking traits. Unlike candidate gene studies, this methodology requires no *a priori* gene knowledge before testing – all, or nearly all, genes are genotyped (Li *et al.*, 2003; Bierut *et al.*, 2007; Saccone *et al.*, 2007; Uhl *et al.*, 2007).

Additional genome-oriented improvements are expected as intergenic regions of the genome that are not present in current genomic maps (International Human Genome Sequencing Consortium, 2004), are successfully mapped (*e.g.* Genovese, *et al.*, 2013) and available for analysis.

The GWAS explosion has necessitated parallel advancements in statistical analysis. One major area of advancement is in detection of genetic effects of small size, which will be necessary to sum up to the total genetic effect of any complex trait. In addition, several techniques have been developed as alternatives to Bonferroni multiple test corrections, including weighted False Discovery Rate calculations (Lin & Lee, 2012) and use of Bayesian mixture modeling (Wei, *et al.*, 2010). Higher resolution statistical methods continue to be developed for GWAS such as extensions of meta-analyses (Evangelou & Ioannidis, 2013), Gene by Environment (GxE) techniques (Thomas, 2010), pathway-based approaches (Wang, *et al.*, 2010), and an assortment newer techniques that show promise in future genetic analyses (Lange, *et al.*, 2014). With the appropriate statistical techniques and the number of large data sets available it is now possible to conduct meta-analyses of multiple studies to highlight the most stable associations for further investigation.

3. Recent meta-analyses of tobacco-related genetics studies

The use of meta-analyses of GWAS data has proven critical to the overall scrutiny of GWAS results. While hypothesis-free genetic association studies can shed new and unexpected light on an area of research, they may be more subject to false positive findings because of the many statistical tests conducted in each GWAS. Where GWAS provide a broad view, meta-analyses allow a focus on those associations most likely to be relevant. Thus, it is important to conduct both qualitative reviews (see for example Picciotto & Kenny, 2013; Schlaepfer *et al.*, 2008; Ray *et al.*, 2010) and quantitative reviews (*i.e.* meta-analyses) as additional empirical research is published. In our review of the literature we found no fewer than five meta-analyses with similar findings that have been conducted on the underlying genetics of nicotine addiction since 2010.

3.1 Results from the A5-A3-B4 gene cluster

In three papers published simultaneously, meta-analyses by Liu *et al.*, (2010), Furberg *et al.*, (2010), and Thorgeirsson *et al.*, (2010) found highly significant associations between two SNP's (rs16969968 in *CHRNA5* and rs1051730 in *CHRNA3*) and cigarettes smoked per day. It is worth noting that these two SNPs are in high linkage disequilibrium (*e.g.* Liu *et al.*, 2010), which indicates that their results are not fully independent of one another, and teasing apart their specific causal associations with smoking behavior will be difficult. Nevertheless, the consistent findings make it more important to understand the role of these SNPs in affecting tobacco-related behaviors.

Although each of these meta-analyses converged on the same conclusion, each had unique strengths. For example, Furberg *et al.*, (2010) was able to analyze a large pooled sample ($N > 74,000$) and conduct follow-up analyses on candidate SNP's identified by the meta-analysis. The follow-up analyses were conducted on the ENGAGE and Ox-GSK consortia resulting in a sample size for the follow-up GWAS of over 140,000. Thorgeirsson *et al.*,

(2010) followed a similar strategy using a merged data set of two independent studies for replication. Multiple regression models were then conducted to determine if other SNP's in the region around the *A5-A3-B4* cluster could account for additional variance not accounted for by rs1051730. Two additional sites of variation (rs2869046 and rs2036534) with low linkage disequilibrium with rs1051730 showed associations with cigarettes smoked per day after controlling for rs1051730. These results suggest that the three variants are each uniquely associated with cigarettes smoked per day. Lastly, Liu *et al.*, (2010) also conducted follow-up primary analyses using data from the 1000 Genomes project, which offered a five-fold increase in polymorphism density in the *A5-A3-B4* cluster compared to HapMap2 (Frazer *et al.*, 2007). Although other SNPs were identified in these meta-analyses as associated with cigarettes smoked per day, only rs16969968 and rs1051730 showed consistent associations across studies.

In a different approach, a pair of meta-analyses by Munafò and colleagues showed that the combined rs16969968/rs1051730 genotype might have additive effects on cigarettes smoked per day. For example, Ware *et al.*, (2011) replicated the association between rs16969968 ($k = 27$ studies) and rs1051730 ($k = 44$ studies) with cigarettes smoked per day. The aggregate effect size across these studies suggests that this association is equivalent to a one cigarette per day increase in smoking per allele, but as the authors point out, they did not have sufficient power to assess the genetic effect of each SNP conditioned on the other. Estimating conservatively by considering these two SNPs as one effectively, this finding suggests cigarette smokers are smoking on average two additional cigarettes per day if they are homozygous for the risk alleles. Put another way, smokers at highest genetic risk based on these two SNP's expose themselves to over 700 additional cigarettes over the course of a year.

The result from Ware *et al.* was replicated by Mufano *et al.*, (2012) in a meta-analysis of six datasets comprising a total sample of over 12,000 participants. The rs16969968/rs1051730 per-allele effect was also found to be associated with lung cancer at a notably high rate (per-allele OR = 1.31). Based on the analysis by Mufano *et al.*, (2012), possession of one of these risk alleles is associated with a 31% increase in lung cancer risk. Possession of both risk alleles is associated with a 62% increase in lung cancer risk. The sum of these meta-analyses suggests a robust association between the rs16969968/rs1051730 genotype and risk for nicotine addiction and the potential for additional risk due to other SNPs in the area that are not in LD with the two major SNPs. Furthermore, variation in these SNP's may have an additive rather than redundant risk association for nicotine addiction and associated lung cancer.

In light of these robust findings it is important to note that the risk for nicotine addiction resulting from variation in these alleles may vary across ancestry. In another meta-analysis of the *A5-A3-B4* cluster, Chen *et al.*, (2012) stratified the analysis by European, Asian, and African ancestry. The authors utilized a "binning" technique to group together correlated SNP's that may constitute the same nicotine outcome association. In particular, the authors constructed a bin (Bin A) tagged by rs16969968 and rs1051730 (in Europeans) that included a total of 52 SNP's ($r^2 = 0.70$). Results showed a significant cross-population association between Bin A and the dichotomous phenotype light (< 10 cigarettes per day) versus heavy

(> 20 cigarettes per day) smoking. Overall, variation within Bin A was associated with an approximately one-third increase in the likelihood of being a heavy smoker (OR=1.33). However, risk for heavy smoking based on these alleles was lower for the European (OR = 1.31) population compared to the Asian (OR = 1.64) and African (OR = 1.62) populations.

Important genetic work remains to be done by extending the results of meta-analyses to other phenotypes, which will increase our understanding of how SNPs might be affecting traits. One approach is to create an overall genetic risk score (GRS) by combining a small set of SNPs from multiple genes implicated by meta-analyses to be associated with smoking behavior. This method allowed Belsky and colleagues to study a group of SNPs (including the *A5-A4-B4* cluster and *CYP2A6*) and smoking-related phenotypes collected from a longitudinal study started when participants were 11 years old and followed up to the age of 38 (Belsky *et al.*, 2013). Results from their analysis showed no association of the GRS with smoking initiation; however, smokers with higher genetic risk progressed to heavier smoking more quickly than other smokers, and smokers who started during adolescence were more likely to convert to daily smoking and progress to heavy smoking more quickly (Belsky *et al.*, 2013). Thus, extending the data from meta-analyses may broaden the search for genetic effects to multiple sites in the genome, and in this case, reveal important developmental roles for SNPs in tobacco-related behaviors.

3.2 Results from other nicotinic receptor gene studies

The focus on the *A5-A3-B4* cluster is relatively recent, and although other nicotinic receptor subunit genes, such as *CHRNA4* and *CHRNA2*, have not been nominated in meta-analyses, much of the prior neurobiological research on nicotine addiction focused on the high affinity nicotinic $\alpha 4\beta 2$ receptor type (sometimes in conjunction with $\alpha 5$ subunits) (Dani *et al.*, 2011). The well-established and critical role that $\alpha 4\beta 2$ types of receptors play in behavioral responses to nicotine makes it important to mention current genetic findings for their genes.

Recent research by Kamens *et al.*, (2013) on two SNPs in *CHRNA4* (rs1044394 and rs1044396) showed nominal associations with nicotine dependence while no SNP within *CHRNA2* was significant. In a follow-up replication the two *CHRNA4* SNPs showed no association with nicotine dependence within the replication sample. Although the SNPs in *CHRNA4* have not shown up as significant in the meta-analyses, Kamens *et al.*, point out that their findings related to rs1044394 are consistent with two other studies suggesting that this SNP might be worth additional analysis. Specifically, rs1044394 is found in exon 5 of *CHRNA4*, as are two SNPs previously associated with nicotine addiction (rs1044396 and rs1044397) (Feng *et al.*, 2004). SNP rs1044396 was also associated with regular and heavy smoking compared to never-smoker controls (Breitling *et al.*, 2009). Similarly for the gene encoding the $\beta 2$ receptor subunit, *CHRNA2*, SNPs located in and around the gene have not passed the cutoffs for significance in meta-analyses. There have, however, been several positive findings. For example, the *CHRNA2* SNP, rs2072660, has been associated with smoking initiation in women (Greenbaum *et al.*, 2006), while smoking cessation was associated with a highly correlated SNP, rs2072661 (Conti *et al.*, 2008, Perkins *et al.*, 2009). Finally, rs2072658, which is in the promoter region of *CHRNA2*, has been associated with subjective effects of smoking (Ehringer *et al.*, 2007; Hoft *et al.*, 2011).

3.3 Results from the *CYP2A6* gene

Efforts to study tobacco-related behaviors have included the genetics of nicotine metabolism. The *CYP2A6* gene encodes an enzyme that is the major metabolizer of nicotine (Benowitz & Jacob, 1994; Nakajima *et al.*, 1996; Messina *et al.*, 1997; reviewed by Hukkanen *et al.*, 2005). Several studies have implicated functional allelic variation in *CYP2A6* with differences in nicotine metabolism, as well as risk for nicotine addiction (*e.g.* Koudsi *et al.*, 2010; Mwenifumbo & Tyndale, 2009; Malaiyandi *et al.*, 2005). An earlier meta-analysis of *CYP2A6* studies (Carter, *et al.*, 2004) showed no consistent association with smoking status or amount smoked. However, the more recent meta-analysis using samples from the ENGAGE consortium (Thorgeirsson *et al.*, 2010) found a relationship between SNP's in this region and cigarettes smoked per day. Follow-up replication of these effects in the Ox-GSK data confirmed these findings. Further support for the importance of *CYP2A6* was revealed in research on the joint effects of the *A5-A3-B4* cluster and *CYP2A6*, which showed independent and additive effects of allelic risk for the two chromosomal regions on cigarettes smoked per day and nicotine dependence (Wassenaar, *et al.*, 2011). Additional support for the role of *CYP2A6* in nicotine addiction can be found in studies that examine nicotine metabolite ratio (NMR) as a proxy for *CYP2A6* allelic variation. For example, using a within-person design, Sofuoglu *et al.*, (2012) found high nicotine metabolizers reported greater cravings and higher dependency following overnight abstinence as well as greater dose-response relationship between intravenous nicotine administration and elevated cardiac response, as well as greater self-reported reward (feeling “high”, “stimulated”, “good effects”).

Through the studies described above that utilize larger sample populations and by combinations of studies into meta-analyses, the genes for metabolizing nicotine and some nicotine receptor subunits have become strong candidates to explain some of the variance in tobacco-related behaviors. The strength of these results also makes the case to find independent support through other research approaches to provide biological explanations for these genetic associations.

4. Support for the genetic findings from independent lines of research

Meta-analyses have demonstrated the importance of genetic variation in the *A5-A3-B4* and the *CYP2A6* genes in tobacco-related behaviors, but other types of biological studies can provide clues as to the functional mechanisms that underlie the association. One approach is to analyze the tissues in which these genes are expressed. The *A5-A3-B4* cluster of genes encode nicotine receptor subunits that are expressed in many tissues, but the primary sites of interest are several brain regions that regulate nicotine-related behaviors. Work on these regions is described in section 4.1. The *CYP2A6* gene is expressed most abundantly in the liver, and findings on this tissue are presented in section 4.2. Making the full connection of tobacco-related behaviors to function at the level of the whole tissue, the gene, and the allele, will be difficult but is critical for translation of this basic research into treatment.

4.1 Brain regions expressing the *A5-A3-B4* gene cluster and behavioral responses to nicotine

Given the genetic results for the *A5-A3-B4* cluster of genes, recent tests for SNP function have focused on those nicotinic receptors containing $\alpha 5$, $\alpha 3$, and $\beta 4$ protein subunits and the regions of the brain in which receptors containing these subunits are expressed. The receptors are present in several brain regions including the pineal gland, thalamus, and olfactory bulb (see Baldwin *et al.*, 2011) as well as in the ventral tegmental area, substantia nigra, hippocampus, and cortex (see Tuesta *et al.*, 2011 for a review). However, only the habenula (primarily the medial habenula; MHb) and its primary output nucleus, the interpeduncular nucleus (IPN), show the presence of all three receptor subunit types, and at relatively high levels compared to their expression in the rest of the brain (Boulter *et al.*, 1990; Marks *et al.*, 1998; Sheffield, *et al.*, 2000; Salas *et al.*, 2003 & 2004, Tuesta *et al.*, 2011). It is important to note that the exact combinations of subunits into functional nicotinic receptors in MHb and IPN neurons are not known. It also remains to be established whether the expression of these genes is coordinated (*e.g.* Jackson *et al.*, 2013). Furthermore, these subunits likely occur in receptors that include the other α and β subunits (reviewed by Fowler *et al.*, 2008), making the results very difficult to interpret. Nevertheless, the coincidence of genetic studies that implicate the *A5-A3-B4* cluster genes, and their expression in brain regions related to nicotine's psychoactive effects, has caused intense examination of the brain regions.

An important means to tease out the functional role of these receptor subunit genes in the neurobiology of nicotine addiction is through genetically engineered mice (Stoker & Markou, 2013; Drenen & Lester, 2012). This work is in early stages and the results from different groups are sometimes conflicting suggesting more work needs to be done. It also emphasizes how challenging this work is despite the background of robust genetic associations. Recent research by Gallego *et al.* (2012) provides an illustration of the relationship between expression of the *A5-A3-B4* cluster genes, the presence of nicotinic receptors containing $\alpha 5$, $\alpha 3$, and/or $\beta 4$ subunits, and nicotine withdrawal, a behavior associated with addiction. In their paper, Gallego and colleagues describe their development of transgenic mice that over-express the *A5-A3-B4* cluster (called TgCHRNA5/A3/B4). Although these mice did not appear to have higher expression of the $\alpha 5$ subunit, they did have higher concentrations of $\alpha 3\beta 4$ -containing nicotinic receptors in the MHb and possibly the IPN, as well as greater expression of $\beta 4$ -containing receptors in the CA1 region of the hippocampus, compared to their wild-type littermates. Behaviorally, the TgCHRNA5/A3/B4 mice showed greater preference for, and sensitivity to, nicotine compared to the wild-type mice. In contrast, research by Frahm *et al.*, (2011) suggests *CHRNA4* over-expression (in mice known as Tabac mice) alters nicotinic receptor activity in the MHb, and leads to nicotine aversion in two-bottle choice and forced drinking tests.

Additional studies in mice have focused on variation in $\alpha 5$ -containing nicotinic receptors due to the central role that receptors with this subunit play in nicotine addiction phenotypes (primarily aversion). As mentioned above, the SNP rs16969968 found within the *CHRNA5* gene alters function of receptors containing this subunit, and is associated with smoking-related behaviors in humans (Bierut *et al.*, 2008). This SNP is also known as D398N,

referring to the fact that the mutant allele replaces an aspartate (D) with an asparagine (N) in the $\alpha 5$ subunit. In the case of this SNP, over-expression of the $\alpha 5$ subunit with the allele associated with nicotine dependence (N) in the Tabac mice, which already over-express $\beta 4$, results in reduced aversion to nicotine in comparison to controls that did not receive the $\alpha 5$ subunit gene manipulation (Frahm *et al.*, 2011). It will be important to test a comparison group expressing $\alpha 5$ with the D allele, so that the allelic effects can be compared in mice that are equivalent except in which $\alpha 5$ allele they possess.

In contrast, Fowler and colleagues report that $\alpha 5$ knockout mice increase their nicotine intake, and that re-expression of $\alpha 5$ in these knockout mice brings the mice back to levels of intake equivalent to wild-type mice (Fowler *et al.*, 2011). It is difficult to compare the results of Gallego *et al.*, Frahm *et al.*, and Fowler *et al.*, due to the differences in engineering needed to create the mice, and the different measures used to assess nicotine-related behaviors. Despite these differences, their results highlight the fact that meta-analyses seem to correctly place importance on these regions in the genome to tobacco-related behaviors.

Additional assays performed *in vitro* have also contributed to the understanding of the functional implications of these alleles without the confounds and expense inherent to the use of transgenic mice. Flora *et al.* found that alleles of SNPs located between *CHRNA4* and *CHRNA3* altered mRNA expression in neuron-like and lung cancer-like cell lines. They also showed DNA containing these SNPs bind to nuclear proteins in an allele-specific fashion, which suggests a functional consequence of the alleles at these SNPs. In a different study, Gallego *et al.*, (2013) showed that alleles at rs1948, which is located within the 3'-untranslated region of the *CHRNA4* gene, alter expression levels of a reporter gene. Finally, Wang *et al.*, (2009 and 2013) showed that several SNPs located near *CHRNA5* are related to expression of the gene as measured in post-mortem brain samples, and are associated with risk of nicotine dependence, independent of the effects of the D398N variation described above. These types of studies support the genetic association tests by providing potential mechanisms by which the SNPs exert their effect.

4.2 Functional studies of *CYP2A6*

The *CYP2A6* gene has several variants that result in an enzyme with a diminished (*CYP2A6**9 or *CYP2A6**12 alleles) or completely blocked function (*CYP2A6**2 or *CYP2A6**4 alleles) that have been studied for associations with tobacco-related phenotypes (Benowitz *et al.*, 2006; Huang *et al.*, 2005; O'Loughlin *et al.*, 2004; Karp *et al.*, 2006; Audrain-McGovern *et al.*, 2007; Wassenaar *et al.*, 2011). But these reports were not always consistent, and SNPs in or near *CYP2A6* did not remain significant in meta-analysis until the work on cigarettes per day (Thorgeirsson *et al.*, 2010). Despite this limited meta-analytic support, there are multiple lines of evidence to suggest a relationship between *CYP2A6* and tobacco-related behaviors.

As mentioned above, there are genetic variants in the human population that prevent metabolism of nicotine by the *CYP2A6* enzyme. The drug methoxsalen mimics these variants by blocking the enzyme pharmacologically. Methoxsalen has been shown to diminish smoking and the desire to smoke (Sellers *et al.*, 2000 & 2003). It is thought that these effects are due to methoxsalen's ability to modulate nicotine metabolism in the liver

(Benowitz & Jacob, 1994; Nakajima *et al.*, 1996; Messina *et al.*, 1997), but it is important to note that *CYP2A6* might be expressed in the brain at levels high enough to affect nicotine action in close proximity to the nicotinic receptors that regulate behavior (Ferguson & Tyndale, 2011).

The approach of utilizing mouse research to buttress human studies for *CYP2A6* is complicated by the fact that the *CYP* family of genes are highly similar to one another, so targeted engineering of a specific gene is very difficult. Despite the difficulties, genetic manipulation of the *Cyp2a5* gene (the mouse orthologue of human *CYP2A6*) has occurred. Mice lacking *Cyp2a5* metabolize nicotine at less than 10% the rate of wild-type mice, so that equivalent blood levels of nicotine can be achieved with smaller doses (Li *et al.*, 2013). Importantly, these mice also display reward-related behaviors (such as Conditioned Place Preference) to concentrations of nicotine 4-fold lower than wild-type mice (Li *et al.*, 2013). This result is consistent with humans carrying mutations in *CYP2A6* that cause low rates of nicotine metabolism and less smoking (Sellers *et al.*, 2000). It is important to note that these mice, known as *Cyp2a(4/5)bgs*-null mice, lack both *Cyp2a5* and *Cyp2a4*, as well as all of the *Cyp2b* genes. While *Cyp2a4* appears to play no role in nicotine metabolism, a small role for the *Cyp2b* genes was demonstrated in these mice (Li *et al.*, 2013). This minimal role for genes other than *Cyp2a5* does not significantly alter the conclusion concerning the importance of *Cyp2a5* in modulating the rewarding properties of nicotine in mice.

The results described above raise a question concerning the SNP rs4105144 near *CYP2A6* that was significantly associated with cigarettes smoked per day in a meta-analysis (Thorgeirsson *et al.*, 2010). Specifically, are the alleles at rs4105144 functional and able to alter nicotine metabolism rates by altering the activity of *CYP2A6*? Differing rates of nicotine metabolism would provide a potential mechanistic explanation for the genetic findings. Unfortunately, the potential for rs4105144 to be related directly to nicotine metabolism is complicated based on our limited state of knowledge. Specifically, this SNP is not in the protein-coding parts of the gene, nor in any region shown to regulate the gene's expression. Based on current knowledge, Thorgeirsson and colleagues suggested that perhaps rs4105144 was in linkage disequilibrium with another unknown functional SNP within the gene. However, as shown in section 5.2, the ENCODE database suggests another potential explanation.

5. The ENCODE project as an aid to predicting function for tobacco behavior-related SNPs

Although replication of genetic studies will be needed to assure the value of pursuing therapeutic goals based on the research, replication does not lead to testable hypotheses of the biological mechanism by which alleles at a particular SNP cause a genetic association. This gap between genetic association and biological causation is due in large part to the lack of knowledge concerning allelic function. In other words, there is rarely any information to explain why two alleles might affect an individual differently, even if the SNP is located in or near a nicotinic receptor subunit gene.

A large step was taken to bridge this information gap by the ENCODE project (The ENCODE Project Consortium, 2012; see also approximately 30 papers published simultaneously with the Consortium's paper and found at nature.com/ENCODE). This work generated a database that can serve as a hypothesis-generating tool to assess functional causes for genetic associations (Maurano, *et al.*, 2012; Schaub *et al.*, 2012). Publication of the ENCODE data represented a significant first step toward the overall goal of the project to understand the biochemical function of every nucleotide in the human genome. The major areas of function tested in the ENCODE publication answer the following questions at near single-base resolution: 1) Is the nucleotide transcribed? 2) Is the nucleotide part of a transcription factor binding site (TFBS)? 3) Is the nucleotide part of a DNase I Hypersensitive Site (DHS)? 4) Is the nucleotide part of a region with altered chromatin marks (histone modifications or DNA methylation)? And finally, 5) Does the nucleotide physically interact with DNA at great distance from it on the chromosome? These questions are shown schematically in Fig. 1.

Whether all of the measures tested in ENCODE should be labeled as functional is open to debate, and the debate has been significant (Doolittle, 2013; Eddy, 2013; Graur *et al.*, 2013). For instance, if a nucleotide is transcribed as part of an RNA molecule, it is not clear whether changes to that nucleotide would alter the RNA's function or any other molecular function. However, transcription (or any of the other measures) is at the very least, a mark that can be assigned to the nucleotide, and can guide researchers to testable hypotheses concerning function.

Another issue, beyond that of assigning function, is the range of scope of the ENCODE experiments. Although the study generated enormous volumes of data, the data are limited by two parameters. First, only a subset of all the possible cultured cell types was tested. Even fewer tissues were tested, and this tissue-limitation is a particular problem with respect to the brain. Concern that one's favorite target tissue is not included in the ENCODE data might be tempered by the fact that the project is in an early phase and has laid groundwork for adding results from future studies. Second, the set of transcription factors and chromatin modifications chosen to study is limited, but a comprehensive study of all transcription factors and chromatin modifications is not yet practical. An additional concern is that perhaps not all functions that cause genetic associations (*e.g.* mRNA stability) would be considered genomic functions as envisioned in ENCODE, and perhaps will not be reported by the ENCODE project.

It is not our intent to over-emphasize the limitations of the ENCODE data or proclaim that it will solve all the mysteries of allelic function. ENCODE is a work in progress and new experimental data from additional cell types, transcription factors and perhaps novel assays will expand the project in the future. In addition, each field of study will have its own unique needs that must be addressed through experimentation and addition of new data to ENCODE. For example, data from brain tissue is needed for genetic studies of human behavior. Cultured cells of neuronal origin might provide some relevant data, but will not capture all of the information related to the regions and sub-regions (*e.g.* IPN and MHb) of the brain.

An important point that might diminish some of these concerns comes from analysis of ENCODE data on a genome-wide basis. Both Maurano *et al.* (2012) and Schaub *et al.* (2012) found that functional SNPs are enriched in the list of SNPs that have been associated in GWAS reports when compared to SNPs that have never been nominated. The enrichment increases as the confidence of the GWAS reports increases, so that in comparing SNPs that have been nominated once, versus nominated and replicated within a report, versus nominated with independent replication, there is an increase in the proportion of SNPs with known or potential function (Maurano *et al.* 2012). Not surprisingly, the SNP-function relationship is also strongest when the functional data come from multiple sources of information (Schaub *et al.*, 2012). These results suggest that there is a large pool of SNPs with potential function waiting to be analyzed.

Thus, despite some limitations, ENCODE is making a massive amount of functionally relevant data accessible to the genetics community. A significant result of the increased accessibility is the potential to examine GWAS-nominated SNPs for co-localization with the four biochemical marks tested in ENCODE.

5.1 Using ENCODE to assess allele function from the literature

The explanatory power provided by ENCODE data for function of a SNP nominated by a GWAS can vary in strength based on the nature of the information already present in the literature. A case provided by Maurano *et al.*, (2012) examines SNP rs385893, which had been associated with blood platelet count in a previous GWAS (Soranzo, *et al.*, 2009), but which had no known function at the point of Soranzo's publication. In discussing their genetic results, Soranzo and colleagues pointed out that this SNP is 400 kb (thousand base pairs) from *JAK2*, a gene known to play a key role in maturation of certain white blood cells. Importantly, it was already known that this gene is mutated in approximately one-half of individuals with thrombocytosis, a disorder that causes destruction of platelets (Wickrema and Crispino, 2007). *JAK2* was mentioned by Soranzo and colleagues due to this potential relationship to platelet count, but there were no data directly supporting a relationship between the SNP and *JAK2*. In fact, *JAK2* is not the closest gene to rs385893, that honor goes to *AK3*, which is within 10 kb of the SNP. *AK3*, however, has no known relationship to platelet count, and thus the genetic association between rs385893 and platelet count had no strong candidate gene, much less function, to support the GWAS results. The potential for *JAK2* to be a causative gene in platelet count increased considerably when data from ENCODE showed that the DNA around rs385893 physically interacts with the *JAK2* promoter (Maurano *et al.*, 2012). It remains to be demonstrated experimentally if or how the genotype at rs385893 alters this physical interaction with the *JAK2* promoter, but if the interaction is important for expression of *JAK2*, then it would provide a link between the SNP and platelet count. Thus, this finding provides a testable hypothesis concerning differences in *JAK2* promoter activity based on alleles at the SNP, and demonstrates the value of ENCODE to bridge the gap between genetic associations and function.

5.2 Using ENCODE to reveal potential function for a SNP near *CYP2A6*

Although the example above demonstrates the ability to link SNPs identified in genetic associations to functions related to the phenotype being studied, it begs the question of

whether anything can be gained relevant to nicotine and pharmacogenomics. The following example is pertinent to the topic of this review; however, the link to function is not as strong as the example above, and we stress that the results are not yet supported by independent experiments.

This case is based on the SNP rs4105144, which was described previously as being associated with the number of cigarettes smoked per day (Thorgeirsson *et al.*, 2010). This SNP is approximately 2,700 nucleotides upstream of the *CYP2A6* gene, a candidate gene strongly linked to nicotine metabolism and smoking behaviors. In the absence of any data to point to a specific function of the SNP, the proximity of rs4105144 to *CYP2A6* suggested to Thorgeirsson and his colleagues that the genetic association was due to linkage disequilibrium of rs4105144 with SNPs within the gene (Thorgeirsson *et al.*, 2010).

There is a second potential explanation for the association that is not mutually exclusive of the linkage hypothesis proposed by Thorgeirsson *et al.*, (2010). This explanation is revealed by using ENCODE to search for information related to rs4105144 at this website.

The search is started by going to the home page of ENCODE (<http://www.encodeproject.org>) and clicking on the link labeled “Visualization,” which brings up a page with a random site in the genome from the Human Genome Browser. This browser is the major portal to human genomic information, and ENCODE is incorporated within it. The second step is to go to the site for the SNP of interest by clicking on the white box containing the phrase, “enter position, gene symbol or search terms,” and entering rs4105144, and then clicking the button labeled “go.” On the subsequent page, several options appear, the first of which, “NHGRI Catalog of Published Genome-Wide Association Studies,” indicates that this SNP is present in the NHGRI's GWAS database. Not all SNPs associated with phenotypes have been placed in this database, but this option takes one to the correct site in the genome as does the second choice, “Simple Nucleotide Polymorphisms (dbSNP build 130).” Clicking on either of these two options leads to a webpage that shows rs4105144 and a small part of the genome flanking it. Figure 2A presents a portion of the web page, the remainder of which is not shown, but will be described in Section 5.3.

This page's title (UCSC Genome... see red arrow #1 in Fig. 2A) shows the version of the genome on which the ENCODE data are placed. Below the title is a line of navigation buttons (red arrow #2) that allow the viewer to move along the chromosome. The third line (red arrow #3) contains the chromosome number as well as the genomic coordinate, or nucleotide position, of the SNP. In this case the SNP is on chromosome 19 approximately 46 million nucleotides from the tip of the short arm of the chromosome. The fourth line (red arrow #4) is a schematic of the chromosome known as an ideogram showing bands representing Giemsa staining of the DNA. A small vertical red line on the ideogram indicates the approximate position of the SNP on the ideogram.

The center of the page (red arrow #5) is a white box with light blue vertical lines known as the Genome Browser Window that is the working part of the page. The information in this window is organized into tracks, and many of the items in the window are clickable links to

additional information. Having used rs4105144 as the target, the window presents 501 base pairs (bp) of the genome centered on this SNP (shown in more detail in Fig. 2B). Unfortunately for SNP-centered investigations, the SNP's position is shown at the bottom of the browser window as a hash mark beside the SNP name in reverse text (white letters on a green background) so it is necessary to locate the position of the SNP at the bottom of the window before working from the top through the remaining information in the window.

The top-most track in the window shown in Fig. 2B, labeled "Scale," presents a bar 200 bp in length. The second track labeled, "chr 19" gives the nucleotide coordinates of the DNA in the current browser window.

The next track shows any gene identified in this part of the chromosome. The thin line with arrowheads indicates that a gene called *FKSG58* is present in this region and that rs4105144 resides within the intron of the gene. Zooming out 10x with the button in the second line from the top of the page reveals the beginning of the *CYP2A6* gene, also within the *FKSG58* gene (data not shown). While it is not unheard of for one gene to contain another gene within one of its introns, it is rare and suggests further analysis of *FKSG58* before going further. By clicking on the gene name or the line with arrowheads representing the gene, a new window opens. The data from the resulting window shows that *FKSG58* is likely to be an artifact, perhaps from a rare, aberrant cDNA generated in a high-throughput sequencing project. This "gene" serves as a caution to the reader not to accept data as accurate even when present in this reference website for genomic data. *FKSG58* is somewhat unusual in several ways that lead one to suspect that it is an experimental artifact. The gene encodes an 833 nucleotide mRNA, from two exons separated by approximately 318,000 base pairs, and does not include a protein coding region based on current methods of identifying them from RNA sequence data. In addition, there are no known orthologs (genes related by descent) in any other species. Finally, the gene is labeled as, "not on the current assembly," of the human genome (<http://www.ncbi.nlm.nih.gov/gene/100996488>), even though it shows up in the version of the Browser containing the ENCODE data. Based on the sum of the evidence, it is likely safe to conclude that *FKSG58* is an artifact and can be ignored.

Returning to the webpage for the SNP (Fig. 2B), the track for genes is followed by data tracks labeled Transcription, Layered H3K4Me1, Layered H3K4Me3, DNase Clusters, Txn Factor ChIP, SNPS and finally RepeatMasker. These last two tracks are not specific to the ENCODE project and will not be described further.

The track labeled, "Transcription," shows little evidence that this region is transcribed and polyadenylated (i.e. produces an mRNA) in any of the cell types tested by the ENCODE consortium. If the region were highly transcribed the track would be filled with a colored peak or peaks with each color representing a different cell type, the names of which can be revealed by clicking on the track.

The next track, Layered H3K4Me1, contains information on the presence of histones with a specific pattern of chemical modification. The name of the track is based on the name of the histone (H3), the site of modification (lysine at the 4th amino acid position – K4), and the presence of 1 methyl group (Me1) on the lysine. This type of methylation of histone H3 is

associated with enhancers and promoters of genes that are transcribed. Given the lack of a peak in this region there is little likelihood of high concentrations of histones with this type of modification.

The next track, Layered H3K4Me3, shows data for the presence of histone H3 with 3 methyl groups on the lysine at its fourth amino acid. Histones with three methyl groups are found near the start site of transcription of a gene. Although some future ENCODE experiment might reveal a high level of modified histones in this region of DNA, and thus suggest a nearby gene is expressed, there is no current evidence for the presence of this type of histone modification.

The DNase Clusters track shows a light gray bar over the position of rs4105144, which is evidence supporting a direct functional explanation for why this SNP is associated with cigarettes smoked per day. The region indicated by the bar is known as a DNase I Hypersensitive Site (DHS), which is defined by a degree of general openness or accessibility of the DNA sequence to the proteins needed to carry out gene expression. DHS are frequently found in and around genes that are expressed. The test for accessibility involves incubating nuclei isolated from cells with an enzyme known as DNase I, which binds to open DNA and cleaves it. After isolating the fragments of cleaved DNA, the ends of the fragments are sequenced and the sequence is aligned to the genome. Regions of the genome at which many fragments align indicate preferential cleavage by DNase I and thus open DNA. The presence of rs4105144 within a DNase I hypersensitive site suggests that this SNP may be located in a region of altered DNA structure, and that the alleles of this SNP might alter expression of a nearby gene, or genes.

Before going any further with the demonstration, it is important to learn more about how to interpret the information presented on this track. By clicking on the gray box in the DNase Clusters track, a new page called, "DNase Clusters item details," opens up with more information about this particular DHS. From this new page, one learns that the intensity of gray corresponds to a score of 142 out of 1,000, which is weak evidence of cleavage by DNase I. It is, however, worth consideration when combined with the results from the next track described below. It is noteworthy that this page also indicates that this site was identified in one experiment in a single cell type known as H7-hESC, which is an undifferentiated embryonic stem cell. None of the genes nearby this SNP is expressed in this cell type. Therefore, the data on the hypersensitive site might be stronger if tested in a cell type that expressed any of the genes near rs4105144. In sum, this DNase Cluster information points toward future experiments, and should not be taken as conclusive.

The Txn Factor ChIP track presents data that support this argument for a functional role for rs4105144. The dark gray bar in this track shows that one or more transcription (Txn) factors bind to this region of DNA. The data are generated by a Chromatin Immuno-Precipitation (ChIP) method, in which DNA and proteins are chemically cross-linked inside cells, the cells lysed, and the DNA gently sheared. Regions of DNA that are cross-linked to the proteins are protected from shearing and can be purified by precipitation with antibodies against the cross-linked protein, in this case a transcription factor. The many fragments of DNA that were precipitated are then unlinked from the proteins and sequenced. After

alignment of the resulting sequences to the genome, the frequency with which a given fragment of DNA is present in the sequence data is calculated and presented in this track. Clicking on the dark gray box in the track opens a new page called, “ENCODE Transcription Factor ChIP-seq (GR)” (not shown). The page indicates that the transcription factor binding to the DNA around rs4105144 was GR, or Glucocorticoid Receptor (gene name *NR3C1*) and the score was 529 out of 1,000. Although this result was from one experiment and must be replicated, there is supporting evidence in the literature that the *CYP2A6* gene is regulated by the glucocorticoid receptor (Onica *et al.*, 2008).

In the study by Ornica, a DNA region related to the *CYP2A6*'s responsiveness to GR was limited to a small area near the gene's start of transcription. This region is approximately 2,700 nucleotides away from rs4105144, but it is not clear from the report whether any experiment included this SNP. Thus it is an open question as to whether the region around rs4105144 plays a role in the binding of GR to this DNA, and controlling expression of *CYP2A6*. We wish to point out that the relationship between SNP and glucocorticoid receptor is still a putative one that needs to be experimentally verified, for example by testing whether the two alleles of rs4105144 alter GR binding to this region, and then testing for altered expression of *CYP2A6* depending on which allele of rs4105144 is present.

The hypothesis that alleles at this SNP alter binding of the Glucocorticoid Receptor and provide a functional explanation for the SNP's association with a smoking-related trait would not have been contemplated without the ENCODE data. Furthermore, experimental techniques to test this hypothesis are well developed and straightforward to interpret. As more data are added to the ENCODE project, it becomes more likely that additional SNP's associated with nicotine addiction can be tied to genomic function. As analysis of function becomes more common, it may be increasingly useful to develop proficiency in how to navigate the ENCODE page. Additional details on the ENCODE page, and its fit into the Human Genome website, can be found by following the tutorial and the user's guide, which is accessible by going to the Genome Browser home page (<http://genome.ucsc.edu/>), clicking on “Training,” in the left-hand column, and then selecting the link labeled “Genome Browser User's Guide.”

5.3. Additional tracks of ENCODE data not discussed in this review

Fig. 2A presents the default window of data on DNA elements related to ENCODE, but a wide array of additional data tracks are available to add to the browser window. The dark blue bar at the bottom of Fig. 2A labeled “Mapping and Sequencing Tracks,” is just one of 15 sets of tracks that can be added to the browser window. For those interested in methylation of DNA, some of the tracks listed under “Regulation” will provide details of methylation at CpG islands. “Phenotype and Disease Associations” might be of particular interest to readers concerned with genetic association studies, while “Comparative Genomics,” “Neandertal Assembly and Analysis,” and “Denisova Assembly and Analysis,” provides the opportunity to examine sequence conservation across species.

6. Conclusion

An understanding of the relationships between genetics and genomic function is critical if we are to understand the biological mechanisms by which nicotine exerts its effects on behavior, and to use that knowledge to reduce the harm caused by tobacco. Determining which genetic associations are important for further study will be aided by determining if variation in DNA sequence has consequences for genomic function. Analyzing ENCODE data for SNPs nominated by genetic studies of tobacco-related behaviors will prove useful in this effort. There have been many calls in the human behavioral genetics field to find endophenotypes for complex behaviors that are “closer” to the gene (*e.g.* McGuffin *et al.*, 1987). While the data generated by ENCODE do not fit the accepted limitation of endophenotypes to measures obtained from living human beings (Gottesman and Gould, 2003), they do represent a potential avenue to understanding nicotine pharmacogenomics at a molecular level—ENCODE is about as close to the gene as it gets.

Acknowledgments

The authors would like to thank H. Harrington Cleveland, Kerry Hair, Will Horton, Helen Kamens, Marta Byrska-Bishop, and the members of the Vandenbergh lab for fruitful discussion and work that led to the ideas put forward in this work. This work was supported in part by a grant from the National Institute on Drug Abuse (R01DA030389).

GLOSSARY

Allele	One of the possible forms of a genetic variation at a given site in the genome. The type of form can range from the specific base sequence (<i>e.g.</i> a C allele at a site in DNA sequence that has either C or T as the two forms) to the phenotypic outcome of the base sequence (<i>e.g.</i> an albino allele at a site that can generate either a colored or albino coat).
Base and base pair	see nucleotide, with which base is often used interchangeably.
CpG Island	a region of DNA in which the frequency of the DNA sequence CG is elevated above expected. This sequence is the site at which the vast majority of methylated C's are found and are thought to be important for regulation of openness of DNA and gene expression.
DNase I	an enzyme that cleaves DNA without bias toward the sequence being cut, and is used to distinguish regions of DNA that are relatively free of the proteins used to pack the DNA into chromosomes.
Enhancer	A short DNA sequence that promotes expression of a gene in part by allowing transcription factors to bind. An enhancer may be within the gene, closely flanking the gene, or at a great distance.
Heritability	The degree to which a phenotype or trait can be attributed to genes passed down from one generation to the next.

Linkage disequilibrium	when alleles from two separate sites in the genome are inherited together at a rate higher than expected by chance, indicating the alleles occur on one of a chromosome pair (i.e. either the maternally or paternally inherited chromosome). This situation is similar to linkage between two polymorphisms, one of which is non-functional but co-inherited with a second, causative (functional) polymorphism. The opposite situation, linkage equilibrium, reflects the co-inheritance of two alleles at a frequency expected by chance. Equilibrium is similar to no linkage between two polymorphisms.
Nucleotide	a fundamental building block of DNA, consisting of a nitrogen-containing base, a five-carbon sugar, and a phosphate. The base determines the genetic information of a given site in DNA.
rs#	Reference Sequence, or RefSeq, number, which is uniquely assigned by when a SNP is deposited in the dbSNP database.
SNP (Single Nucleotide Polymorphism)	A site in a DNA sequence at which there is variation of one base or nucleotide (<i>e.g.</i> A in place of G). The convention, only loosely held, is that the frequency of the less common base must be >1% for a variant to be considered a SNP.
Transcription Factor	A protein needed to regulate the level of transcription (expression) of a gene. These proteins act by binding to specific sites within and around the gene being regulated. The distance between the site at which transcription starts and the transcription factor binds can be very different from gene to gene.

References

- Audrain-McGovern J, Al Koudsi N, Rodriguez D, Wileyto EP, Shields PG, Tyndale RF. The role of CYP2A6 in the emergence of nicotine dependence in adolescents. *Pediatrics*. Jan; 2007 119(1):e264–74. Epub 2006 Nov 27. PubMed PMID: 17130279. [PubMed: 17130279]
- Baldwin PR, Alanis R, Salas R. The role of the habenula in nicotine addiction. *J Addict Res Ther*. 2011; S1:002. [PubMed: 22493758]
- Belsky DW, Moffitt TE, Baker TB, Biddle AK, Evans JP, Harrington H, Houts R, Meier M, Sugden K, Williams B, Poulton R, Caspi A. Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA Psychiatry*. May; 2013 70(5):534–42. doi: 10.1001/jamapsychiatry.2013.736. PubMed PMID: 23536134; PubMed Central PMCID: PMC3644004. [PubMed: 23536134]
- Benowitz NL, Jacob P 3rd. Metabolism of nicotine to cotinine studied by a dual stable isotope method. *Clin Pharmacol Ther*. 1994; 56(5):483–493. [PubMed: 7955812]. [PubMed: 7955812]
- Benowitz NL, Swan GE, Jacob P 3rd, Lessov-Schlaggar CN, Tyndale RF. CYP2A6 genotype and the metabolism and disposition kinetics of nicotine. *Clin Pharmacol Ther*. Nov; 2006 80(5):457–67. PubMed PMID: 17112802. [PubMed: 17112802]
- Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry*. 2008; 13:368–73. [PubMed: 18227835]
- Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L, Fugman D, Goate AM, Hinrichs AL, Konvicka K, Martin NG, Montgomery

- GW, Saccone NL, Saccone SF, Wang JC, Chase GA, Rice JP, Ballinger DG. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet.* 2007; 16:24–35. [PubMed: 17158188]
- Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X, Saccone NL, Saccone SF, Bertelsen S, Fox L, Horton WJ, Breslau N, Budde J, Cloninger CR, Dick DM, Foroud T, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Kuperman S, Madden PA, Mayo K, Nurnberger J Jr, Pomerleau O, Porjesz B, Reyes O, Schuckit M, Swan G, Tischfield JA, Edenberg HJ, Rice JP, Goate AM. Variants in nicotinic receptors and risk for nicotine dependence. *Am J Psychiatry.* 2008; 165:1163–71. [PubMed: 18519524]
- Boulter J, O'Shea-Greenfield A, Duvoisin RM, Connolly JG, Wada E, Jensen A, Gardner PD, Ballivet M, Deneris ES, McKinnon D, et al. Alpha 3, alpha 5, and beta 4: three members of the rat neuronal nicotinic acetylcholine receptor-related gene family form a gene cluster. *J Biol Chem.* Mar 15; 1990 265(8):4472–82. PMID: 1689727. [PubMed: 1689727]
- Breitling LP, Dahmen N, Mittelstrass K, Rujescu D, Gallinat J, Fehr C, Giegling I, Lamina C, Illig T, Müller H, Raum E, Rothenbacher D, Wichmann HE, Brenner H, Winterer G. Association of nicotinic acetylcholine receptor subunit alpha 4 polymorphisms with nicotine dependence in 5500 Germans. *Pharmacogenomics J.* 2009; 9:219–24. [PubMed: 19290018]
- CDC. The Health Consequences of Smoking: A Report of the Surgeon General. Department of Health and Human Services, Centers for Disease Control and Prevention; National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; Atlanta: 2004. 2004
- Chen LS, Saccone NL, Culverhouse RC, Bracci PM, Chen CH, Dueker N, Bierut LJ. Smoking and genetic risk variation across populations of European, Asian, and African American ancestry – A meta-analysis of chromosome 15q25. *Genet Epidemiol.* 2012; 36:340–351. [PubMed: 22539395]
- Comings DE, Ferry L, Bradshaw-Robinson S, Burchette R, Chiu C, Muhleman D. The dopamine D2 receptor (DRD2) gene: a genetic risk factor in smoking. *Pharmacogenetics.* 1996; 6:73–9. [PubMed: 8845863]
- Conti DV, Lee W, Li D, Liu J, Van Den Berg D, Thomas PD, Bergen AW, Swan GE, Tyndale RF, Benowitz NL, Lerman C. The Pharmacogenetics of Nicotine Addiction and Treatment Consortium. Nicotinic acetylcholine receptor beta2 subunit gene implicated in a systems-based candidate gene study of smoking cessation. *Hum Mol Genet.* 2008; 17:2834–2848. [PubMed: 18593715]
- Dani JA, Jenson D, Broussard, De Biasi M. Neurophysiology of nicotine addiction. *J Addict Res Ther.* 2011; S1:001. [PubMed: 22454789]
- Doolittle WF. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A.* 2013; 110:5294–300. [PubMed: 23479647]
- Eddy SR. The ENCODE project: missteps overshadowing a success. *Curr Biol.* 2013; 23(7):R259–61. [PubMed: 23578867]
- Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet.* 2013; 14(6):379–89. doi: 10.1038/nrg3472. [PubMed: 23657481]
- Feng Y, Niu T, Xing H, Xu X, Chen C, Peng S, Wang L, Laird N, Xu X. A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. *Am J Hum Genet.* 2004; 75:112–21. [PubMed: 15154117]
- Ferguson CS, Tyndale RF. Cytochrome P450 enzymes in the brain: emerging evidence of biological significance. *Trends Pharmacol Sci.* 2011; 32(12):708–14. doi: 10.1016/j.tips.2011.08.005. [PubMed: 21975165]
- Flora AV, Zambrano CA, Gallego X, Miyamoto JH, Johnson KA, Cowan KA, Stitzel JA, Ehringer MA. Functional characterization of SNPs in CHRNA3/B4 intergenic region associated with drug behaviors. *Brain Res.* Sep 5.2013 1529:1–15. doi: 10.1016/j.brainres.2013.07.017. PubMed PMID: 23872218; PubMed Central PMCID: PMC3839343. [PubMed: 23872218]
- Fowler CD, Arends MA, Kenny PJ. Subtypes of nicotinic acetylcholine receptors in nicotine reward, dependence, and withdrawal: evidence from genetically modified mice. *Behav Pharmacol.* 2008; 19:461–84. [PubMed: 18690103]
- Fowler C, Lu Q, Johnson PM, Marks MJ, Kenny PJ. Habenular $\alpha 5$ nicotinic receptor subunit signaling controls nicotine intake. *Nature.* 2011; 471:597–601. [PubMed: 21278726]

- Frahm S, Slimak MA, Ferrarese L, Santos-Torres J, Antolin-Fontes B, Auer S, Ibañez-Tallon I. Aversion to nicotine is regulated by the balanced activity of $\beta 4$ and $\alpha 5$ nicotinic receptor subunits in the medial habenula. *Neuron*. 2011; 70:522–535. [PubMed: 21555077]
- Frazer KA. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–61. [PubMed: 17943122]
- Furberg H. Tobacco and Genetics Consortium. Genome-wide meta-analysis identify multiple loci associated with smoking behavior. *Nature Genet*. 2010; 42:441–447. [PubMed: 20418890]
- Gallego X, Molas S, Amador-Arjona A, Marks MJ, Robles N, Murtra P, Dierssen M. Overexpression of the *CHRNA5/A3/B4* genomic cluster in mice increases the sensitivity to nicotine and modifies its reinforcing effects. *Amino Acids*. 2012; 43:897–909. [PubMed: 22101982]
- Giovino GA, Mirza SA, Samet JM, Gupta PC, Jarvis MJ, Bhala N, Peto R, Zatonski W, Hsia J, Morton J, Palipudi KM, Asma S, GATS Collaborative Group. Tobacco use in 3 billion individuals from 16 countries: an analysis of nationally representative cross-sectional household surveys. *Lancet*. 2012; 380:668–79. [PubMed: 22901888]
- Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry*. 2003; 160:636–45. [PubMed: 12668349]
- Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*. 2013; 5(3):578–90. [PubMed: 23431001]
- Greenbaum L, Kanyas K, Karni O, Merbl Y, Olender T, Horowitz A, Yakir A, Lancet D, Ben-Asher E, Lerer B. Why do young women smoke? I. Direct and interactive effects of environment, psychological characteristics and nicotinic cholinergic receptor genes. *Mol Psychiatry*. 2006; 11:312–322. 223. [PubMed: 16314871]
- Hoft NR, Stitzel JA, Hutchison KE, Ehringer MA. *CHRNA2* promoter region: association with subjective effects to nicotine and gene expression differences. *Genes Brain Behav*. Mar; 2011 10(2):176–85. doi: 10.1111/j.1601-183X.2010.00650.x. PubMed PMID: 20854418; PubMed Central PMCID: PMC3403289. [PubMed: 20854418]
- Huang S, Cook DG, Hinks LJ, Chen XH, Ye S, Gilg JA, Jarvis MJ, Whincup PH, Day IN. *CYP2A6*, *MAOA*, *DBH*, *DRD4*, and *5HT2A* genotypes, smoking behaviour and cotinine levels in 1518 UK adolescents. *Pharmacogenet Genomics*. Dec; 2005 15(12):839–50. [PubMed: 16272956]
- Hukkanen J, Jacob P 3rd, Benowitz NL. Metabolism and disposition kinetics of nicotine. *Pharmacol Rev*. 2005; 57:79–115. [PubMed: 15734728]
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431:931–945. [PubMed: 15496913]
- Kamens HM, Corley RP, McQueen MB, Stallings MC, Hopfer CJ, Crowley TJ, Ehringer MA. Nominal association with *CHRNA4* variants and nicotine dependence. *Genes Brain Behav*. 2013; 12:297–304. [PubMed: 23350800]
- Jackson KJ, Sanjakdar SS, Muldoon PP, McIntosh JM, Damaj MI. The $\alpha 3\beta 4^*$ nicotinic acetylcholine receptor subtype mediates nicotine reward and physical nicotine withdrawal signs independently of the $\alpha 5$ subunit in the mouse. *Neuropharmacology*. 2013; 70:228–35. doi: 10.1016/j.neuropharm.2013.01.017. [PubMed: 23416040]
- Karp I, O’Loughlin J, Hanley J, Tyndale RF, Paradis G. Risk factors for tobacco dependence in adolescent smokers. *Tob Control*. 2006; 15:199–204. [PubMed: 16728750]
- Koudsi NA, Hoffman EB, Assadzadeh A, Tyndale RF. Hepatic *CYP2A6* levels and nicotine metabolism: Impact of genetic, physiological, environmental, and epigenetic factors. *Eur J Clin Pharmacol*. 2010; 66:239–251. [PubMed: 20012030]
- Lange K, Papp JC, Sisnheimer JS, Sobel EM. Next-generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application*. 2014 advanced online publication. doi: 10.1146/annurev-statistics-022513-115638.
- Li L, Jia K, Zhou X, McCallum SE, Hough LB, Ding X. Impact of nicotine metabolism on nicotine’s pharmacological effects and behavioral responses: insights from a *Cyp2a(4/5)bgs*-null mouse. *J Pharmacol Exp Ther*. 2013; 347(3):746–54. doi: 10.1124/jpet.113.208256. PubMed PMID: 24045421; PubMed Central PMCID: PMC3836308. [PubMed: 24045421]

- Li MD, Ma JZ, Cheng R, Dupont RT, Williams NJ, Crews KM, Payne TJ, Elston RC. Framingham Heart Study. A genome-wide scan to identify loci for smoking rate in the Framingham Heart Study population. *BMC Genet.* 2003; 31(4 Suppl 1):S103. [PubMed: 14975171]
- Lin WY, Lee WC. Discovering joint associations between disease and gene pairs with a novel similarity test. *BMC Genet.* 2010; 11:86. doi: 10.1186/1471-2156-11-86. [PubMed: 20920333]
- Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Marchini J. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet.* 2010; 42:436–440. [PubMed: 20418889]
- Marks MJ, Smith KW, Collins AC. Differential agonist inhibition identifies multiple epibatidine binding sites in mouse brain. *J Pharmacol Exp Ther.* 1998; 285(1):377–386. [PubMed: 9536034]
- Malaiyandi V, Sellers EM, Tyndale RF. Implications of CYP2A6 genetic variation for smoking behaviors and nicotine dependence. *Clin Pharmacol Ther.* 2005; 77:145–158. [PubMed: 15735609]
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutayavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012; 337:1190–5. [PubMed: 22955828]
- McGuffin P, Farmer A, Gottesman II. Is there really a split in schizophrenia? The genetic evidence. *Br J Psychiatry.* 1987; 150:581–92. [PubMed: 3307978]
- Messina ES, Tyndale RF, Sellers EM. A major role for CYP2A6 in nicotine C-oxidation by human liver microsomes. *J Pharmacol Exp Ther.* Sep; 1997 282(3):1608–14. PubMed PMID: 9316878. [PubMed: 9316878]
- Munafò MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennen P, Smith GD. Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J Natl Cancer Inst.* 2012; 104:740–748. [PubMed: 22534784]
- Nakajima M, Yamamoto T, Nunoya K, Yokoi T, Nagashima K, Inoue K, Funae Y, Shimada N, Kamataki T, Kuroiwa Y. Role of human cytochrome P4502A6 in C-oxidation of nicotine. *Drug Metab Dispos.* 1996; 24(11):1212–7. PubMed PMID: 8937855. [PubMed: 8937855]
- O'Loughlin J, Paradis G, Kim W, DiFranza J, Meshefedian G, McMillan-Davey E, Wong S, Hanley J, Tyndale RF. Genetically decreased CYP2A6 and the risk of tobacco dependence: a prospective study of novice smokers. *Tob Control.* Dec; 2004 13(4):422–8. [PubMed: 15564629]
- Onica T, Nichols K, Larin M, Ng L, Maslen A, Dvorak Z, Pascucci JM, Vilarem MJ, Maurel P, Kirby GM. Dexamethasone-mediated up-regulation of human CYP2A6 involves the glucocorticoid receptor and increased binding of hepatic nuclear factor 4 alpha to the proximal promoter. *Mol Pharmacol.* 2008; 73:451–60. [PubMed: 17978169]
- Picciotto MR, Kenny PJ. Molecular mechanisms underlying behaviors related to nicotine addiction. *Cold Spring Harb Perspect Med.* 2013; 3:1–12.
- Ray, R.; Schnoll, R.; Lerman, C. Nicotine dependence and pharmacogenetics. In: Steptoe, A., editor. *Handbook of behavioral medicine.* Springer Science+Business Media; Berlin, Germany: 2010.
- Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, Swan GE, Goate AM, Rutter J, Bertelsen S, Fox L, Fugman D, Martin NG, Montgomery GW, Wang JC, Ballinger DG, Rice JP, Bierut LJ. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum. Mol. Genet.* 2007; 16:36–49. [PubMed: 17135278]
- Salas R, Orr-Urtreger A, Broide R, Beaudet AL, Paylor R, De Biasi M. The nicotinic acetylcholine receptor subunit alpha 5 mediates acute effects of nicotine in vivo. *Mol Pharmacol.* 2003; 63:1059–1066. [PubMed: 12695534]
- Salas R, Pieri F. De Biasi M Decreased signs of nicotine withdrawal in mice null for the $\alpha 4$ nicotinic acetylcholine receptor subunit. *J Neurosci.* 2004; 24:10035–10039. [PubMed: 15537871]
- Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 2012; 22(9):1748–59. [PubMed: 22955986]

- Schlaepfer IR, Hoft NR, Ehringer MA. The genetic components of alcohol and nicotine co-addiction: From genes to behavior. *Curr Drug Abuse Rev.* 2008; 1:124–134. [PubMed: 19492010]
- Sellers EM, Kaplan HL, Tyndale RF. Inhibition of cytochrome P450 2A6 increases nicotine's oral bioavailability and decreases smoking. *Clin Pharmacol Ther.* Jul; 2000 68(1):35–43. PubMed PMID: 10945314. [PubMed: 10945314]
- Sellers EM, Tyndale RF, Fernandes LC. Decreasing smoking behaviour and risk through CYP2A6 inhibition. *Drug Discov Today.* Jun 1; 2003 8(11):487–93. Review. PubMed PMID: 12818518. [PubMed: 12818518]
- Sheffield EB, Quick MW, Lester RA. Nicotinic acetylcholine receptor subunit mRNA expression and channel function in medial habenula neurons. *Neuropharmacology.* 2000; 39(13):2591–603. [PubMed: 11044729]
- Sofuoglu M, Herman AI, Nadim H, Jatlow P. Rapid nicotine clearance is associated with greater reward and heart rate increases from intravenous nicotine. *Neuropsychopharmacol.* 2012; 37:1509–1516.
- Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A, Willenborg C, Wright B, Chen L, Li M, Salo P, Voight BF, Burns P, Laskowski RA, Xue Y, Menzel S, Altshuler D, Bradley JR, Bumpstead S, Burnett MS, Devaney J, Döring A, Elosua R, Epstein SE, Erber W, Falchi M, Garner SF, Ghorri MJ, Goodall AH, Gwilliam R, Hakonarson HH, Hall AS, Hammond N, Hengstenberg C, Illig T, König IR, Knouff CW, McPherson R, Melander O, Mooser V, Nauck M, Nieminen MS, O'Donnell CJ, Peltonen L, Potter SC, Prokisch H, Rader DJ, Rice CM, Roberts R, Salomaa V, Sambrook J, Schreiber S, Schunkert H, Schwartz SM, Serbanovic-Canic J, Sinisalo J, Siscovick DS, Stark K, Surakka I, Stephens J, Thompson JR, Völker U, Völzke H, Watkins NA, Wells GA, Wichmann HE, Van Heel DA, Tyler-Smith C, Thein SL, Kathiresan S, Perola M, Reilly MP, Stewart AF, Erdmann J, Samani NJ, Meisinger C, Greinacher A, Deloukas P, Ouwehand WH, Gieger C. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet.* 2009; 41(11):1182–90. doi: 10.1038/ng.467. PubMed PMID: 19820697; PubMed Central PMCID: PMC3108459. [PubMed: 19820697]
- Stoker AK, Markou A. Unraveling the neurobiology of nicotine dependence using genetically engineered mice. *Curr Opin Neurobiol.* 2013; 23:1–7. [PubMed: 23265962]
- Thomas D. Gene--environment-wide association studies: emerging approaches. *Nat Rev Genet.* 2010; 11(4):259–72. doi: 10.1038/nrg2764. [PubMed: 20212493]
- Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsäter A, Flex A, Aben KK, de Vegt F, Mulders PF, Isla D, Vidal MJ, Asin L, Saez B, Murillo L, Blondal T, Kolbeinnsson H, Stefansson JG, Hansdottir I, Runarsdottir V, Pola R, Lindblad B, van Rij AM, Dieplinger B, Haltmayer M, Mayordomo JJ, Kiemeny LA, Matthiasson SE, Oskarsson H, Tyrfinsson T, Gudbjartsson DF, Gulcher JR, Jonsson S, Thorsteinsdottir U, Kong A, Stefansson K. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature.* 2008; 452:638–42. [PubMed: 18385739]
- Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Stefansson K. Sequence variants at *CHRNA3-CHRNA6* and *CYP2A6* affect smoking behavior. *Nature Genet.* 2010; 42:448–453. [PubMed: 20418888]
- Tuesta LM, Fowler CD, Kenny PJ. Recent advances in understanding nicotinic receptor signaling mechanisms that regulate drug self-administration behavior. *Biochem Pharmacol.* 2011; 82:984–995. [PubMed: 21740894]
- Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. *BMC Genet.* 2007; 8:10. [PubMed: 17407593]
- Vandenbergh DJ, O'Connor RJ, Grant MD, Jefferson AL, Vogler GP, Strasser AA, Kozlowski LT. Dopamine receptor genes (*DRD2*, *DRD3* and *DRD4*) and gene-gene interactions associated with smoking-related behaviors. *Addict Biol.* 2007; 12:106–16. [PubMed: 17407504]

- Wang JC, Cruchaga C, Saccone NL, Bertelsen S, Liu P, Budde JP, Duan W, Fox L, Grucza RA, Kern J, Mayo K, Reyes O, Rice J, Saccone SF, Spiegel N, Steinbach JH, Stitzel JA, Anderson MW, You M, Stevens VL, Bierut LJ, Goate AM, COGEND collaborators and GELCC collaborators. Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in *CHRNA5*. *Hum Mol Genet*. 2009; 18(16):3125–35. [PubMed: 19443489]
- Wang JC, Spiegel N, Bertelsen S, Le N, McKenna N, Budde JP, Harari O, Kapoor M, Brooks A, Hancock D, Tischfield J, Foroud T, Bierut LJ, Steinbach JH, Edenberg HJ, Traynor BJ, Goate AM. Cis-Regulatory Variants Affect *CHRNA5* mRNA Expression in Populations of African and European Ancestry. *PLoS One*. 2013; 8(11):e80204. doi: 10.1371/journal.pone.0080204. [PubMed: 24303001]
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010; 11(12):843–54. doi: 10.1038/nrg2884. [PubMed: 21085203]
- Ware JJ, van den Bree MBM, Munafò MR. Association of the *CHRNA5-A3-B4* gene cluster with heaviness of smoking: A meta-analysis. *Nicotine Tob Res*. 2011; 13:1167–1175. [PubMed: 22071378]
- Wassenaar CA, Dong Q, Wei Q, Amos CI, Spitz MR, Tyndale RF. Relationship between *CYP2A6* and *CHRNA5-CHRNA3-CHRNA4* variation and smoking behaviors and lung cancer risk. *J Natl Cancer Inst*. 2011; 103:1342–1346. [PubMed: 21747048]
- Wei YC, Wen SH, Chen PC, Wang CH, Hsiao CK. A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies. *Eur J Hum Genet*. 2010; 18(8):942–7. doi: 10.1038/ejhg.2010.51. [PubMed: 20407469]
- WHO report on the global tobacco epidemic. warning about the dangers of tobacco. World Health Organization; Geneva: 2011.
- Wickrema A, Crispino JD. Erythroid and megakaryocytic transformation. *Oncogene*. 2007; 26:6803–6815. PubMed: 17934487. [PubMed: 17934487]

Highlights

- Covers recent GWAS meta-analyses on nicotine addiction
- Reviews research on brain regions associated with the meta-analysis findings
- Describes the ENCODE database and application to genetic association studies
- Explores examples of SNPs with potential genomic function revealed by ENCODE

Fig. 1.

The genomic marks tested by the ENCODE consortium are shown in this image of DNA from the ENCODE project website. The gray boxes indicate the techniques used to detect the marks. From left to right: 5C detects long-distance interactions of two regions of DNA mediated by proteins, often transcription factors (not described in this review); DNase-seq and FAIRE-seq detect DNase Hypersensitivity; CHIP-seq detects transcription factor binding; Computational predictions and RT-PCR, and RNA-Seq detect transcription; and finally, methylation is indicated by the CH₃ tag on the DNA. The CH₃CO represents hydroxymethylation, which is related to methylation, but also not covered in this review. Credit: Darryl Leja (NHGRI) and Ian Dunham (EBI).

Fig. 2.

The Genome Browser window for ENCODE data related to SNP rs4105144. **a)** The first page that appears provides a view with a resolution of 501 base pairs around the SNP in the Browser Window (central white box with vertical light blue lines). Active (clickable) text and drop-down menus will manipulate the resolution in base pairs, the position along the chromosome, and the types of data shown in the Browser Window (red numbered arrows, see text for description). Note that the position of the SNP is indicated at the bottom of the Browser Window with tracks of data labeled on the left (*e.g.* Transcription). Options for additional tracks of data lie below the part of the webpage shown here and are described in section 5.3 of the text. **b)** A high-resolution view of the browser window in (a) was created by choosing, “View,” from the top right of the bar in (a) and choosing the PDF/PS option. Following the position of rs4105144 at the bottom of the image and looking up, the SNP resides within a transcription factor binding site or sites (darker gray bar in the Txn Factor ChIP track), to the left of a DNase Hypersensitive Site (light gray bar in the DNase Clusters track). Clicking on the Txn Factor ChIP track reveals that GR (Glucocorticoid Receptor) binds to this region. See text for details.