Genome-Wide Meta-Analysis Identifies Regions on 7p21 (AHR) and 15q24 (CYP1A2) As Determinants of Habitual Caffeine Consumption

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:41246960

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Genome-Wide Meta-Analysis Identifies Regions on 7p21 (AHR) and 15q24 (CYP1A2) As Determinants of Habitual Caffeine Consumption

Marilyn C. Cornelis1–9, Keri L. Monda2–9, Kai Yu3–9, Nina Paynter4–9, Elizabeth M. Azzato3, Sirií N. Bennett5, Sonja I. Berndt3, Eric Boerwinkle6, Stephen Chanock3, Nilanjan Chatterjee3, David Couper7, Gary Curhan8, Gerardo Heiss3, Frank B. Hu1, David J. Hunter1, Kevin Jacobs9, Majken K. Jensen1, Peter Kraft9, Maria Teresa Landi7, Jennifer A. Nettleton6, Mark P. Purdie3, Preetha Rajaraman3, Eric B. Rimm1, Lynda M. Rose4, Nathaniel Rothman3, Debra Silverman3, Rachael Stolzenberg-Solomon3, Amy Subar3, Meredith Yeager3, Daniel I. Chasman4,*; Rob M. van Dam10,*; Neil E. Caporaso3,*

Abstract
We report the first genome-wide association study of habitual caffeine intake. We included 47,341 individuals of European descent based on five population-based studies within the United States. In a meta-analysis adjusted for age, sex, smoking, and eigenvectors of population variation, two loci achieved genome-wide significance: 7p21 (P = 5.2 × 10−10), near AHR, and 15q24 (P = 5.2 × 10−10), between CYP1A1 and CYP1A2. Both the AHR and CYP1A2 genes are biologically plausible candidates as CYP1A2 metabolizes caffeine and AHR regulates CYP1A2.

Citation: Cornelis MC, Monda KL, Yu K, Paynter N, Azzato EM, et al. (2011) Genome-Wide Meta-Analysis Identifies Regions on 7p21 (AHR) and 15q24 (CYP1A2) As Determinants of Habitual Caffeine Consumption. PLoS Genet 7(4): e1002033. doi:10.1371/journal.pgen.1002033

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received November 17, 2010; Accepted February 6, 2011; Published April 7, 2011

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: ARIC is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022, R01HL076951, R01HL09367, and R01HL086949; National Human Genome Research Institute contract U01HG004622; and NIH contract HHSN26820062526C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the NIH and NIH Roadmap for Medical Research. The NHS Breast Cancer GW scan was performed as part of the Cancer Genetic Markers of Susceptibility initiative. The authors have declared that no competing interests exist.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: caporasni@mail.nih.gov (NEC); ephrmvd@nus.edu.sg (RMvD); dchasman@rics.bwh.harvard.edu (DIC)

These authors contributed equally to this work.

These authors were joint senior authors on this work.
Introduction

Caffeine (1,3,7-trimethylxanthine) is the most widely consumed psychoactive substance in the world. Although demographic and social factors have been linked to habitual caffeine consumption, twin studies report a large heritable component. Through a comprehensive search of the human genome involving over 40,000 participants, we discovered two loci associated with habitual caffeine consumption: the first near AHR and the second between CYP1A1 and CYP1A2. Both the AHR and CYP1A2 genes are biologically plausible candidates, as CYP1A2 metabolizes caffeine and AHR regulates CYP1A2. Caffeine intake has been associated with manifold physiologic effects and both detrimental and beneficial health outcomes. Knowledge of the genetic determinants of caffeine intake may provide insight into underlying mechanisms and may provide ways to study the potential health effects of caffeine more comprehensively.

Author Summary

Caffeine is the most widely consumed psychoactive substance in the world. Although demographic and social factors have been linked to habitual caffeine consumption, twin studies report a large heritable component. Through a comprehensive search of the human genome involving over 40,000 participants, we discovered two loci associated with habitual caffeine consumption: the first near AHR and the second between CYP1A1 and CYP1A2. Both the AHR and CYP1A2 genes are biologically plausible candidates, as CYP1A2 metabolizes caffeine and AHR regulates CYP1A2. Caffeine intake has been associated with manifold physiologic effects and both detrimental and beneficial health outcomes. Knowledge of the genetic determinants of caffeine intake may provide insight into underlying mechanisms and may provide ways to study the potential health effects of caffeine more comprehensively.

Results

We performed a meta-analysis of 47,341 individuals of European descent, derived from five studies within the US, the Atherosclerosis Risk in Communities (ARIC, N = 8,945) Study, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO, N = 4,942), the Nurses' Health Study (NHS, N = 6,774), the Health Professionals Follow-Up Study (HPFS, N = 4,023), and the Women's Genome Health Study (WGHS, N = 22,658). Sample characteristics are presented in Table 1. Caffeine intake was assessed using semi-quantitative food frequency questionnaires (FFQ) that included questions on the consumption of caffeinated coffee, tea, soft drinks, and chocolate.

Study-level genomic inflation factors (λ) were low ranging from 1.00 (PLCO) to 1.03 (HPFS), suggesting that population stratification was well controlled (Figure S1). A total of 433,781 imputed and genotyped SNPs passed our stringent criteria for the meta-analysis. Test statistic inflation at the meta-analysis level revealed no evidence of notable underlying population substructure (λ = 1.04, Figure 1).

Two loci reached genome-wide significance with no evidence for significant between-study heterogeneity (Table 2, Figure 2 and Figure 3, Table S1). The strongest associated SNP (rs4410790, \( P = 2.4 \times 10^{-19} \)). The second strongest associated SNP (rs2470893, \( P = 5.2 \times 10^{-14} \)) mapped to 15q24 within the bidirectional promoter of the CYP1A1-CYP1A2 locus [6,7]. A synonymous coding SNP (rs2472304, \( P = 2.5 \times 10^{-7} \)) in CYP1A2 exon 7 that was highly correlated with 6 other SNPs but not correlated with rs2470893 (\( r^2 = 0.18 \), HapMap CEU) was amongst the highest ranked loci in our meta-analysis (Table 2). Although we only considered variants that were imputed with high probability,
we also conducted a sensitivity analysis restricting our sampling to individuals with genotyped data (Table 2). Regression coefficients remained essentially unchanged, but P-values were less significant reflecting the reduced sample size (rs4410790: $P = 4.0 \times 10^{-18}$; rs2470893 $P = 9.5 \times 10^{-18}$). Similar results were also observed when men and women were examined separately (Table S2). Had the analysis been performed instead by discovery at genome-wide significance ($P < 5 \times 10^{-8}$) in the WGHS followed by replication in meta-analysis of the remaining cohorts, only SNPs at the same loci would have met Bonferroni corrected standards of significance. In a post-hoc investigation of study heterogeneity in which we compared WGHS to the remaining studies combined, there was significant heterogeneity for rs4410790 ($P = 0.01$), although this could be attributable to chance.

Based on the well-established biological link between smoking and AHR [8], and CYP1A2 [9] and caffeine consumption behavior [2], we explored the role of cigarette smoking (Table 3). Compared to our primary model that adjusted for smoking, a model not adjusted for smoking yielded slightly attenuated associations and when restricting analyses to ‘never smokers’ similar regression coefficients were observed as for the complete study population. These findings suggest that smoking is unlikely the cause of the associations observed in our GWAS of caffeine intake.

We further conducted 21 candidate gene analyses and found significant gene-based associations (Bonferroni corrected for the total number of human genes) between CYP2C9 ($P = 0.023$), and ADORA2A ($P = 0.011$) and caffeine intake in addition to CYP1A2 and AHR (Table 4).

**Discussion**

In the first GWAS of caffeine intake in a total of 47,341 individuals from five U.S. studies, loci at 15q24 and 7p21 achieved genome-wide significance. CYP1A2 at 15q24 and AHR at 7p21 are attractive candidate genes for caffeine intake. At plasma concentrations typical of humans ($<100 \, \mu M$), caffeine is predominantly ($\sim 95\%$ of a dose) metabolized by CYP1A2 via N1-, N3-, and N7-demethylation to its three dimethylxanthines, namely, theobromine, paraxanthine, and theophylline, respectively [10]. CYP1A2 expression and activity vary 10- to 60-fold between individuals [11]. Human CYP1A2 is located immediately adjacent to CYP1A1 in reverse orientation and the two genes share a common 5′-flanking region [12]. At least 15 AHR response elements (AHRE) reside in this bidirectional promoter region and rs2470893 is located in AHRE6 (originally reported as AHRE5[7]) which correlates with transcriptional activation of both CYP1A1 and CYP1A2 [6,7]. CYP1A1 expression in the liver (the target tissue for caffeine metabolism) is low and there is little evidence that this enzyme contributes to caffeine metabolism. This contrasts with the tissue specific expression of CYP1A2 in the liver, which suggests
Table 2. Genome-wide meta-analytic results for caffeine consumption (P < 9 × 10⁻⁶).

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>Closest gene(s) (± 100 kb)</th>
<th>Chr, position</th>
<th>Chr, closest gene(s)</th>
<th>Closest gene(s) (± 100 kb)</th>
<th>Chr, position</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4410790</td>
<td>AHR</td>
<td>7</td>
<td>17,231,912</td>
<td>SCAMP5,PPCDC</td>
<td>15</td>
</tr>
<tr>
<td>rs4729893</td>
<td>CYP1A2</td>
<td>15</td>
<td>72,831,291</td>
<td>CYP1A2, CYP1A1, CYP2C9, CSK, SCAMP5,PPCDC</td>
<td>3</td>
</tr>
<tr>
<td>rs4729893</td>
<td>CYP1A2</td>
<td>15</td>
<td>72,831,291</td>
<td>CYP1A2, CYP1A1, CYP2C9, CSK, SCAMP5,PPCDC</td>
<td>3</td>
</tr>
<tr>
<td>rs6495122</td>
<td>ULK3, SCAMP2, MP1, LMAN1L, CYP1A2, CSK, C14orf17</td>
<td>15</td>
<td>72,831,291</td>
<td>CYP1A2, CYP1A1, CYP2C9, CSK, SCAMP5,PPCDC</td>
<td>3</td>
</tr>
<tr>
<td>rs12148488</td>
<td>ADORA2A</td>
<td>15</td>
<td>73,168,953</td>
<td>ADORA2A</td>
<td>15</td>
</tr>
</tbody>
</table>

Further evidence supporting its role in caffeine metabolism. The observation that a stronger association exists for SNPs upstream of the gene suggests that variation in CYP1A2 gene expression probably affects caffeine intake. The protein product of AHR, AhR, is a ligand-activated transcription factor that, upon binding, partners with ARNT and translocates to the nucleus where it regulates the expression of a number of genes including CYP1A1 and CYP1A2. There is marked variation in AhR binding affinity across populations, but so far no polymorphisms have been identified that account for this variation [13]. The most studied SNP, rs2066853 (R554K), is located in exon 10, a region of AHR that encodes the transactivation domain [13]. Although this SNP was associated with caffeine in the current study (P = 0.0004), our strongest signal mapped upstream of AHR, suggesting variation in AHR expression has a key role in propensity to consume caffeine. An interaction between CYP1A2 and AHR could be biologically plausible; however, we did not find any evidence supporting statistical interaction between the top two loci (data not shown).

Human and animal candidate gene studies for caffeine intake and related traits have focused on various other genes linked to caffeine’s metabolism and targets of action. In our candidate gene analyses, we observed significant gene-based associations between CYP2C9 and ADORA2A and caffeine intake in addition to CYP1A2 and AHR. CYP2C9 catalyzes the N⁷-demethylation and C⁸-hydroxylation of caffeine to theophylline and 1,3,7-trimethyluric acid (a minor metabolite), respectively; but its role relative to CYP1A2 is generally small [10]. In amounts typically consumed from dietary sources, caffeine antagonizes the actions of adenosine at the adenosine A₂A receptor (ADORA2A) [2], which plays an important role in the stimulating and reinforcing properties of caffeine [14,15]. Polymorphisms of ADORA2A have been previously implicated in caffeine-induced anxiety as well as habitual caffeine intake [16,17].

All studies contributing to our GWAS of caffeine intake were US-based. Consistent with the adult caffeine consumption pattern of this country, coffee contributed to well over 80% of caffeine intake. Previous studies suggest that some of the heritability underlying specific caffeine sources (i.e. coffee and tea) may be distinct in relation to total caffeine intake [18]. To evaluate the robustness of findings, we conducted an additional GWAS analysis using caffeinated coffee intake as the outcome variable yielding the same strong signals (rs4410790: 1.4 × 10⁻⁶, rs2470893: 3.6 × 10⁻¹⁹).

Imprecision in phenotypic assessment and differences across studies could have limited the scope of our discovery. Although dietary intake obtained by FFQ is subject to misclassification, validation studies in subsamples of the included studies indicated that the consumption of caffeine-containing beverages is assessed with good accuracy [19,20,21]. The cubic root transformation we applied to reported caffeine intakes, however, limits interpretation of the effect estimates. The crude weighted mean difference in caffeine intake between homozygote genotypes was 44 mg/d for rs4410790 and 38 mg/d for rs2470893 (Table S3 and S4). The two SNPs together, however, explained between 0.06 and 0.72% of the total variation in caffeine intake across studies suggesting additional variants remain to be discovered [22]. Finally, our GWAS assumed an additive genetic model and based on study-level results (Figure 1 and Figure 2) potential non-linear effects will require confirmation in future studies.

Caffeine intake has been associated with pleotropic physiologic effects in relation to both detrimental and beneficial health outcomes [23]. Our current study provides insights into the primary pathways underlying caffeine intake. Knowledge of the genetic determinants of caffeine intake may provide insight into underlying mechanisms and may provide ways to study the
potential health effects of caffeine more comprehensively by using genetic determinants as instrumental variables for caffeine intake or by taking into consideration caffeine-gene interactions. With the exception of nicotine dependency and the associated nicotinic receptor, genes that influence traits associated with dependency have been difficult to identify. The association of caffeine consumption with genes involved in metabolism or its regulation (CYP1A2 and AhR, respectively) illustrates that it is feasible to use GWAS to identify genetic determinants of other behavioral traits that are assessed with lower accuracy. We also recognize that the identified variants could influence regulation of their genomic elements distant from the known, high profile, neighboring candidate genes. In conclusion, we identified two loci related to caffeine consumption that will be worthy of further investigation with regard to both beneficial and toxic effects of caffeine as well as the extensive group of carcinogens, drugs, and xenobiotics also metabolized through action of the regulation of the gene products of CYP1A2 and AHR.

Material and Methods

Ethics Statement
This study was conducted according to the principles expressed in the Declaration of Helsinki. All participants in the contributing studies gave written informed consent including consent for genetic analyses. Local institutional review boards approved study protocols.

Study Populations
We conducted a meta-analysis of 47,341 individuals of European descent, sourced from Atherosclerosis Risk in Communities (ARIC, N = 8,976), the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO, N = 4,942), the Nurses' Health Study (NHS, N = 6,774), the Health Professionals Follow-Up Study (HPFS, N = 4,023), and the Women's Genome Health Study (WGHS, N = 22,658) to identify novel loci associated with habitual caffeine consumption. Study population descriptions and genotyping quality control for data generated with either the Affymetrix 6.0 or the Illumina Infinium arrays (HumanHap300, 550 or 610 arrays) are provided in Text S1 and Table S5 and S6.

Caffeine Intake Assessment
In the NHS, every 2 to 4 years of follow-up diet was assessed using a validated semi-quantitative food frequency questionnaire (FFQ) [24]. For the present analysis, we included the participants' mean caffeine intakes of the 1984 (first year in which caffeinated and decaffeinated coffee were differentiated) and 1986 FFQs. The following caffeine-containing foods and beverages were included in the FFQ: coffee with caffeine, tea, cola and other carbonated
beverages with caffeine, and chocolate. For each item, participants were asked how often, on average, they had consumed a specified amount of each beverage or food over the past year. The participants could choose from nine frequency categories (never, 1–3 per month, 1 per week, 2–4 per week, 5–6 per week, 1 per day, 2–3 per day, 4–5 per day and 6 or more per day). Intakes of nutrients and caffeine were calculated using US Department of Agriculture food composition sources. In these calculations, we assumed that the content of caffeine was 137 mg per cup of coffee, 47 mg per cup of tea, 46 mg per can or bottle of cola or other caffeinated carbonated beverage, and 7 mg per 1 oz serving of chocolate candy. We assessed the total intake of caffeine by summing the caffeine content for the specified amount of each food multiplied by a weight proportional to the frequency of its use. In a validation among a subsample of this cohort, we obtained high correlations between intake of caffeinated coffee and other caffeinated beverages from the FFQ and four 1-week diet records (coffee, r = 0.78; tea, r = 0.93; and caffeinated sodas, r = 0.85) [21].

In the WGHS, caffeine intake was assessed at baseline (1991) using the same FFQ and caffeine algorithm as the NHS [25]. HPFS participants have been followed with repeated FFQs every 4 years. Caffeine-intake was assessed by the same methods as described above for the NHS cohort. In a validation study in a subsample of participants, we obtained high correlations between consumption of coffee and other caffeinated beverages estimated from the FFQ and consumption estimated from repeated 1-wk diet records (coffee: r = 0.83; tea: r = 0.62; low-calorie caffeinated sodas: r = 0.67; and regular caffeinated sodas: r = 0.56)[21]. For the present analysis, we included the participants mean caffeine intakes of the 1986 (baseline) and 1990 FFQs.

In the ARIC study, caffeine consumption was quantified at the baseline (1987–1989) examination from an interview-administered

**Figure 3. Forest plots of the meta-analysis for the two caffeine-associated loci.** A) rs4410790 and B) rs2470893. The contributing effect from each study is shown by a square, with confidence intervals indicated by horizontal lines. The contributing weight of each study to the meta-analysis is indicated by the size of the square. The meta-analysis estimate is shown at the bottom of each graph.
doi:10.1371/journal.pgen.1002033.g003

**Table 3.** Genome-wide meta-analysis of caffeine consumption (P<10^-6): Smoking effects.

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>Chr</th>
<th>EA</th>
<th>Not Adjusted for Smoking</th>
<th>Never Smokers</th>
<th>Current Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>β</td>
<td>P</td>
<td>ρhet*</td>
<td>N</td>
</tr>
<tr>
<td>rs4410790</td>
<td>7</td>
<td>T</td>
<td>−0.15</td>
<td>8.2×10^-18</td>
<td>18</td>
</tr>
<tr>
<td>rs2470893</td>
<td>15</td>
<td>T</td>
<td>0.12</td>
<td>5.0×10^-13</td>
<td>2143</td>
</tr>
<tr>
<td>rs2472304</td>
<td>15</td>
<td>A</td>
<td>0.07</td>
<td>2.4×10^-8</td>
<td>2140</td>
</tr>
<tr>
<td>rs6495122</td>
<td>15</td>
<td>A</td>
<td>−0.07</td>
<td>5.2×10^-6</td>
<td>2143</td>
</tr>
<tr>
<td>rs12148488</td>
<td>15</td>
<td>T</td>
<td>−0.07</td>
<td>1.9×10^-6</td>
<td>2143</td>
</tr>
</tbody>
</table>

Chr, chromosome; EA, effect allele; ρhet*, P value for between study heterogeneity.
doi:10.1371/journal.pgen.1002033.t003
66-item semi-quantitative FFQ[19,20]. The Harvard Nutrition Database was used to assign caffeine (and nutrient) content to each of the food and beverage line items. Line items quantifying consumption of caffeine-containing beverages included sodas (regular and diet), coffee, and tea. The frequency of consumption of each of these items was multiplied by their caffeine content and summed across all beverages to obtain a total caffeine intake value.

Caffeine intake in the PLCO trial was assessed at the randomization phase (between 1992–2001) using responses from a FFQ developed at the National Cancer Institute called the Diet History Questionnaire (DHQ). The DHQ was previously validated against four 24 hour dietary recalls [26] and asks about consumption frequency of 124 food items over the past 12 months, including the primary sources of caffeine: coffee, tea, and soft drinks. For soft drinks, participants selected among 10 possible consumption frequency response categories from “never” to “6 or more cups per day.” For all three of the above beverages, consumption frequency was queried together as cups per unit time ranging from “none”, “1 can or bottle”, “1–2 ounces or 1 can or bottle”, “2–16 ounces or 1 can or bottle” or “>16 ounces or >1 can or bottle.” Frequency and portion size for coffee and tea were queried together as cups per unit time ranging from “never” to “6 or more cups per day.” For all three of the above beverages, participants were asked the proportion of the time each were consumed in decaffeinated form (almost never or never, about half of the time, about ¼ of the time, almost always or always). From these responses daily consumption of caffeine was computed taking into account the caffeine content, portion size, and frequency of intake. Caffeine estimates were derived from two 24-hour dietary recalls administered in the 1994-96 Continuing Survey of Food Intake by Individuals (CSFII)[27], a nationally representative survey conducted during the period when the DHQ was being administered. Individual foods/beverages reported on the recalls were placed in food groups consistent with items on the DHQ, and weighted mean nutrient values based on survey data were derived for adults stratified by sex using methods previously described [28].

### Imputation

Each study used either MACH [29] (ARIC, NHS, HPFS, WGHS) or IMPUTE [30] (PLCO) to impute up to ~2.5 million autosomal SNPs with NCBI build 36 of Phase II HapMap CEU data (release 22) as the reference panel. Genotypes were imputed for SNPs not present in the genome-wide arrays or for those where genotyping had failed to meet the quality control criteria. Imputation results are summarized as an “allele dosage” (a fractional value between 0 and 2), defined as the expected number of copies of the minor allele at that SNP.

### Phenotype Harmonization and Model Selection

The algorithm used for the calculation of caffeine intake was study-specific to allow for differences in questionnaires and consumption habits in different study populations. Raw caffeine-intake measures were skewed across studies and after exploring a variety of transformation options, we found that a cubic-root transformation was very close to the most optimal transformation identified by the Box-Cox procedure and was used to ensure normality of the residuals. Our final models were also adjusted for normality of the residuals. Our final models were also adjusted for...
age (continuous), sex, case-control status (if applicable), study-site (if applicable), smoking status (never, former, and current: 2 categories), and study specific eigenvectors (see Table S5 for study-specific models). Adjustment for smoking status was appropriate given the strong correlation between smoking and caffeine intake that might impede our ability to uncover caffeine-specific loci. Each study collected information on smoking status at the time FFQ were administered. A flexible modeling approach was used to accommodate the different methods by which smoking was collected across studies, but all included never, former and two categories of current smokers. Further adjustments for body-mass-index did not change results appreciably.

Study-Level GWAS
Each study performed genome-wide association testing for normalized caffeine-intake across ~2.5 million SNPs, based on linear regression under an additive genetic model. Analyses were adjusted for additional covariates as described above and further detailed in Table S5. Imputed data (expressed as allele dosage) were examined using ProbABEL [31] or R (scripts developed in-house). The genomic inflation factor λ for each study as well as the meta-analysis was estimated from the median $\chi^2$ statistic.

Meta-Analysis
Meta-analysis was conducted using a fixed effects model and inverse-variance weighting as implemented in METAL (see URLs in Text S1). The software also calculates the genomic control parameter and adjusts each study’s standard errors. Fixed effects analyses are regarded as the most efficient method for discovery in the GWAS setting [32]. Heterogeneity across studies was investigated using the $I^2$ statistic [33]. We applied stringent quality filters to imputed SNPs prior to meta-analysis; removing those with $<0.02$ MAF and/or with low imputation quality scores. The latter was defined as $R^2 \leq 0.80$ for SNPs imputed with MACH and proper_infi $\leq 0.7$ for SNPs imputed with IMPUTE. X and Y chromosome, pseudosomal and mitochondrial SNPs were not included for the present analysis. We retained only SNP-phenotype associations that were based on results from at least 2 of the 10 participating studies and if greater than 50% of the samples contributing to the results were genotyped. Additional checks for experimental biases were implemented for notable associations including manual inspection of SNP (if imputed, an assayed SNP in high LD) cluster plots, and evaluation of HWE, and comparison of study MAFs to the HapMap CEPH panel. We considered P-values $<5 \times 10^{-8}$ to indicate genome-wide significance [34].

Candidate Gene–Based Analyses
We examined 515 SNPs in 23 genes (± 50 kb) either previously studied or members of the key biological pathway: ‘Caffeine metabolism’ (KEGG [35], supplemented with candidates from [10,36]) for association with caffeine consumption in our GWA meta-analysis sample. SNPs mapping to TAS2R10, 43 and 46, implicated in the oral detection of caffeine, did not pass our stringent QC criteria and thus were not included. Gene-based analyses were performed using VEGAS [37]. The software applies a test that incorporates information from a set of markers within a gene (or region) and accounts for LD between markers by using simulations from the multivariate normal distribution. The number of simulations per gene is determined adaptively. In the first stage, 1000 simulations are performed. If the resulting empirical P value is less than 0.1, 10000 simulations are performed. If the empirical P value from 10000 simulations is less than 0.0001, the program will perform 1000000 simulations. At each stage, the simulations are mutually exclusive. For computational reasons, if the empirical P value is 0, then no more simulations will be performed. An empirical P value of 0 from 1000000 simulations can be interpreted as $P<1 \times 10^{-6}$, which exceeds a Bonferroni-corrected threshold of $P<2.8 \times 10^{-6}$ [0.05/17,387 (number of autosomal genes)].

Supporting Information
Figure S1 QQ plots for study-level GWAS of caffeine consumption. Results for genotyped and imputed SNPs denoted by red and blue points, respectively. (TIFF)

Figure S2 Regional association plots of the two caffeine-associated loci. SNPs are plotted with their meta-analysis P-values (as -log10 values) as a function of genomic position (NCBI Build 36). In each panel, the index association SNP is represented by a diamond. Estimated recombination rates (taken from HapMap CEU) are plotted to reflect the local LD structure. SNP color indicates LD with the index SNP according to a scale from $r^2 = 0$ to $r^2 = 1$ based on pairwise $r^2$ values from HapMap CEU. Plots were created using LocusZoom (see URLs). (TIFF)

Table S1 Genome-wide meta-analysis of caffeine consumption: All SNPs $P<10^{-4}$. (DOCX)

Table S2 Genome-wide meta-analysis of caffeine consumption ($P<10^{-6}$): Gender and study effects. (DOCX)

Table S3 Mean caffeine intakes (mg/d) by rs4410790 genotype. (DOCX)

Table S4 Mean caffeine intakes (mg/d) by rs2470893 genotype. (DOCX)

Table S5 Study-specific genotyping, imputation and statistical analysis. (DOCX)

Table S6 Sample quality control. (DOCX)

Text S1 Study population descriptions and URLs. (DOC)

Acknowledgments
ARIC: The authors thank the staff and participants of the ARIC study for their important contributions.

NHS and HPFS: The NHS Breast Cancer GW scan was performed as part of the Cancer Genetic Markers of Susceptibility initiative of the NCI. We particularly acknowledge the contributions of R. Hoover, A. Hutchinson, K. Jacobs and G. Thomas. The NHS/HPFS type 2 diabetes GWAS is a component of a collaborative project that includes 13 other GWAS funded as part of the Gene Environment-Association Studies (GENEVA) under the NIH Genes, Environment and Health Initiative (GEI). Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center and Genotyping Centers (Broad Institute of MIT and Harvard, and Johns Hopkins University Center for Inherited Disease
Author Contributions

Conceived and designed the experiments: NEC RMvD MCC. Analyzed the data: MCC KLM KY NP DIC. Wrote the paper: MCC KLM KY NP DIC RMvD NEC. Genetic data provider: EBR GC FBH PK DJH DIC NEG. Study management: NEC RMvD MCC. Critical review of the manuscript: MCC KLM KY NP EMA SNB SIB EB SC NC DG GC GH FBH DJH KJ MKJ PK MT1 JAN MPP PR EBR LMR NR DS RS-S AS MY DIC RMvD NEC.

References