Supplemental Information

Phenotype Analysis: Asthma vs. COPD

We have performed a comparative analysis to evaluate whether the asthma cohort was phenotypically distinct from the COPD population (Supplemental Tables 9 - 12).

We defined asthma as a positive response to the question “Has a doctor ever told you that you have asthma?” or “Have you ever had asthma?” on any of the available questionnaires derived from the individual studies. The agreement between these two different questions, where available to analyze, was very good, leading us to think that a subject’s report of asthma reflects an actual medical diagnosis (kappa statistic with 95% confidence intervals for individual studies are as follows: ARIC 0.82 CI 0.80-0.85, CARDIA 0.81 CI 0.77-0.83, JHS 0.80 CI 0.78-0.83).

We further refined our asthma sample population by filtering out patients with other chronic lung conditions such as COPD. We defined COPD as a positive response to the question “Has a doctor ever told you that you have chronic bronchitis or COPD?” or “Have you ever had chronic bronchitis or COPD?” on any of the available questionnaires derived from the individual studies.

After filtering out subjects with self-reported COPD, we compared this refined asthma cohort (which was used in our meta-analysis) against those subjects with COPD to determine the effectiveness of this strategy (Supplemental Tables 9 - 12). First, we find that our refined asthma cohort is significantly more allergic than either the control group or the COPD group based on the response to the questions “Has a doctor ever told you that you have hay fever or seasonal allergies?” or “Have you ever had hay fever or seasonal allergies?” (p-value (two-tail): Asthma vs. COPD, p-value 0.0001 for ARIC, CARDIA, MESA; Asthma vs. control, p-value 0.0001 for ARIC, CARDIA, MESA). This does not include subjects from the JHS study, for which no allergy data was available.

Next we found that the number of subjects with low lung function was significantly higher in the COPD group compared to the number of subjects with low lung function in the refined asthma group for most of the studies used in the meta-analysis (p-value (two-

tail): ARIC, p-value 0.0001, CARDIA p-value 0.045, JHS, p-value 0.69, MESA p-value 0.0039). Low lung function was defined as FEV_1 < 70% of predicted based on race and sex specific NHANES III prediction equations adjusting for age and height, or FEV1/FVC less than lower limits of normal for age, race and sex.

Next we found that the number of ever-smokers compared to never-smokers was significantly higher in the COPD group compared to the refined asthma group for most of the parent studies used in the meta-analysis (p-value (two-tail): ARIC, p-value 0.0001, CARDIA p-value = 0.074, JHS, p-value 0.038, MESA p-value 0.0001).

Finally, we found that those subjects in the COPD cohort were on balance significantly older than those subjects in the refined asthma cohort (Median +/- SD, t-test p-value: ARIC, asthma 53.5 yo +/- 5.9, COPD 56.50 yo +/- 5.7, p-value 0.00042; CARDIA, asthma 40 yo +/- 3.63, COPD 40 yo +/- 3.74, P-value = 0.57; JHS, asthma 51 yo +/- 12.6, COPD 55 yo +/- 12.3, p-value = 0.67, MESA, asthma 59 yo +/- 10.2, COPD 67 yo +/- 10.0, p-value < 0.0001). In analysis that was not shown in this manuscript, we stratified subjects by allergy status, smoking status, and lung function, but none of these covariates affected the results and so were not included.

Taken together, these results argue that our method of defining asthma by self-report combined with the exclusion of those with self-reported COPD successfully allowed us to isolate a cohort of asthmatics that was phenotypically distinct from subjects with COPD. The fact that our study overall shows good replication of loci that resulted from studies where asthma was defined by either physician-diagnosis or by physiological criteria further substantiates this claim.

CARe QC

CARe sample handling procedures, QC measures, and data management have been described previously (Lettre et al. 2011). Briefly, samples were genotyped at the Broad Institute using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy6.0) according to the manufacturer's recommendations. A subset of 24 markers including a gender confirmation assay were also genotyped using the Sequenom MassArray System. Genotypes were called using Birdseed v1.33. Quality control steps were performed

using the software PLINK(Purcell et al. 2007), EIGENSTRAT(Price et al. 2006), and PREST-Plus (http://fisher.utstat.toronto.edu/sun/Software/Prest/). Multiple QC procedures were performed including: confirming gentotype concordance between Suquenom iPLEX and Affy6.0; removing samples with a genome-wide genotyping success rate <95%, SNPs with genotyping success rate <90%, monomorphic SNPs, and SNPs that mapped to several genomic locations; removing poor quality DNA (identified by estimating heterozygosity rates); removing sample duplicates, contaminated samples, and cryptic relationships (identified by using genome-wide genotype data to estimate identity-by-descent between all pairwise combinations); outlier samples were removed (identified based on nearest neighbor and "clustering based on missingness" analyses in PLINK); removing SNPs with minor allele frequency (MAF) <1% or with genotyping success rate <95%; in JHS, excluding SNPs with an unusually high number of Mendel errors; and excluding SNPs that showed association with specific chemistry plates. Because several different ethnic groups were represented, with the expectation of differing genotype frequencies and admixture, no filters were applied for Hardy-Weinberg probability values.

CARe Data Management

The institutional review board associated with each CARe cohort have reviewed participation in CARe. The Committee on the Use of Humans as Experimental Subjects of the Massachusetts Institute of Technology has approved analysis within the CARe project. All protected health information has been removed from CARe data and individuals are identified only by unique randomly generated patient identifiers. Genotype and phenotype data from the CARe cohorts has been deposited to dbGAP (Musunuru et al. 2010).

Description of the CARe Cohorts used in the analysis.

CARe cohorts have been described previously and are summarized below (Lettre et al. 2011):

1. Atherosclerosis Risk in Communities (ARIC).

The ARIC study is a prospective population-based study of cardiovascular diseases in 15,792 individuals age 45 to 64 years at the time of initial examination (1987 – 1989), drawn from 4 U.S. communities (suburban Minneapolis, Minnesota; Washington County, Maryland; Forsyth County, North Carolina, and Jackson, Mississippi). Only self-reported African-American participants are included in this analysis. Genotype data was available from 2,989 African American individuals.

2. Coronary Artery Risk Development in young Adults (CARDIA).

The CARDIA study is a prospective, multi-center investigation of the natural history and etiology of cardiovascular disease in 5,115 individuals age 18 to 30 years of age at the time of initial examination (1985 – 1986) and drawn from four communities: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. Each participant's age, race, and sex were self-reported during the recruitment phase and verified during the baseline clinic visit. Genotype data were available on 955 African-American individuals.

3. Jackson Heart Study (JHS).

The Jackson Heart Study is a prospective population-based study to evaluate common complex diseases among 5,301 African Americans age 34 to 84 years at the time of initial examination (2000 – 2004) and drawn from the Jackson, Mississippi metropolitan area. Genotype data were available on 3,030 African-American individuals (some JHS participants are also enrolled in ARIC, and were analyzed with the ARIC dataset – 2,145 individuals are uniquely associated with JHS)

4. Multi-Ethnic Study of Atherosclerosis (MESA).

The Multi-Ethnic Study of Atherosclerosis (MESA) is a study of subclinical cardiovascular disease in a diverse population of 6,814 individuals age 45 to 84 years at the time of initial examination (2000 - 2002) and drawn from six field centers: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California-Los Angeles (Bild 2002). Genotype data were available on 1,636 African-American individuals.

Principal component analysis (PCA)

We used PCA as implemented in EIGENSTRAT (Price et al. 2006) on the CARe African-American Affy6.0 genotype data as described previously. Briefly, two datasets were used as reference populations: 1,178 European Americans derived from a multiple sclerosis GWA study and 756 Nigerians from the Yoruba region derived from a hypertension GWA study. PCA was also used as a screening tool to detect extreme sample outliers before quality control checks.

Genetic association analysis

Prior to genetic association testing, we removed additional samples that had passed all quality control filters described above but would have caused problems in the interpretation of the results. These include: samples with missing gender information (ARIC=85), samples with different IDs that share >90% of the their genome identity-by-descent (IBD)(ARIC=56; JHS=1), samples unlikely to be from African Americans based on principal component analysis results (ARIC=8; CARDIA=2), samples that had a high number of discordant genotypes at SNPs common to both the Affy6.0 platform and the ITMAT-BROAD-CARe (IBC) array (ARIC=3), seven samples from the ARIC dataset that were also present in the JHS dataset based on IBD metrics, and participants who were younger than 18 years old at baseline (CARDIA=5)(Lettre et al. 2011). Thus, the following numbers of African-American participants were available for analysis: ARIC=2,830, CARDIA=949, CFS=521, JHS=2,144, and MESA=1,636 (Total N=8,090). The CFS cohort was not used in meta-analysis because the low number of cases with asthma precluded robust association analysis with logistic regression methods.

Supplemental Table and Figure Legends

Supplemental Table 1: Demographics and exclusion criteria for the control group. Listed in the first row is the starting number of African American subjects without a diagnosis of asthma (subdivided by gender) that were identified from questionnaire data for each of the 4 parent studies used in this analysis. See methods section and Supplemental Table 6 for further details about the questionnaire data that was used to

identify subjects with asthma. Moving down the table, next listed for each study are the exclusion criteria that were applied to arrive at the final n for the control group. Chronic respiratory symptoms were determined from the questionnaire data listed in Supplemental Table 7. Lung disease other than asthma was determined from the questionnaire data listed in Supplemental Table 8. Low lung function was defined as FEV_1 < 70% of predicted based on race and sex specific NHANES III prediction equations adjusting for age and height, or FEV1/FVC less than lower limits of normal for age, race and sex (see material and methods). CARe quality control metrics were defined previously (Lettre et al. 2011). The "total" column under each parent study refers to the total number of subjects positive for the associated exclusion criterion. The ""unique/removed" column under each parent study refers to those subjects who had not been filtered due to the presence of other selection criteria - i.e. this is the number of subjects that were positive only for the one associated selection criterion. The "remaining" column under each parent study refers to the number of subjects remaining after that stage of filtering.

Supplemental Table 2: Demographics and exclusion criteria for the asthma group. Listed in the first row is the starting number of African American subjects with asthma (subdivided by gender) that were identified from questionnaire data for each of the 4 parent studies used in this analysis. See methods section and Supplemental Table 6 for further details about the questionnaire data that was used to identify subjects with asthma. Moving down the table, next listed for each study is the exclusion criteria that were applied to arrive at the final n for the asthma group. Lung disease other than asthma was determined from the questionnaire data listed in Supplemental Table 8. CARe quality control metrics were defined previously (Lettre et al. 2011). The "removed" column under each parent study refers to the total number of subjects positive for the associated exclusion criterion. The "remaining" column under each parent study refers to the number of subjects remaining after that stage of filtering.

Supplemental Table 3: Replication of European asthma associated SNPs in the CARe African-American meta-analyses. Reported SNP refers to the top SNP at the particular locus from the referenced GWA studies, and CARe SNP refers the top SNP in the CARe African American cohort. EAF-CARe is the effect allele frequency for the reported SNP in CARe. EAF is the effect allele frequency for the "better" SNP (i.e. the

SNP with the strongest signal of replication) in CARe that was found in the analysis. DOE (direction of effect): + indicates that the CARe SNP effect was in the same direction as the reported SNP. I/G indicates whether the CARe SNP was imputed (I) or genotyped (G). RSQ, imputation r2 for imputed CARe SNPs. Bold is used to indicate CARe SNPs that either exceeded the Bonferoni corrected significance threshold or were nominally significant and in the same effect direction as the reported SNP.

Supplemental Table 4: Adjustment of meta-analysis results for local ancestry. Reported SNP refers to the top SNP at the particular locus from the referenced GWA studies, and CARe SNP refers the top SNP in the CARe African American cohort. EAF-CARe is the effect allele frequency for the reported SNP in CARe. EAF is the effect allele frequency for the "better" SNP (i.e. the SNP with the strongest signal of replication) in CARe that was found in the analysis. P-value (local ancestry) is the p-value reported in table 2 adjusted for the confounding effects of admixture by using HAPMIX software. Effect direction: + indicates that the CARe SNP effect was in the same direction as the reported SNP. Bold is used to indicate CARe SNPs that exceeded the Bonferoni corrected significance threshold.

Supplemental Table 5: Conditional analysis of signals in the *RAD50/IL13* region and the *IL1RL1/IL18R1* region. Conditioning for the effect of rs17622991 eliminated evidence for association for rs224012. However, conditioning for the effect of rs224012 did not abolish the association signal of rs17622991. Conditional analysis of the signals in *IL1RL1/IL18R1* did not distinguish the signal at the European ancestry SNP from the signal at the African ancestry SNP.

Supplemental Table 6: Questionnaire questions to establish asthma diagnosis. Subjects with asthma were defined based on an affirmative response to any of these study questions and included both physician-diagnosed and self-reported asthma.

Supplemental Table 7: Questionnaire questions to determine history of chronic respiratory symptoms. These symptoms could be consistent with asthma, and so subjects who provided an affirmative answer to these questions were excluded from the control group.

Supplemental Table 8: Questionnaire questions to establish diagnoses consistent with chronic lung disease. Subjects who provided an affirmative answer to these questions were excluded from both the asthma group and the control group.

Supplemental Table 9: Comparison of phenotypic variables between subjects with self-reported asthma or self-reported COPD in ARIC. Key variables that distinguish patients with asthma from patients with COPD (allergic status, lung function, smoking status, and age) were compared in order to determine if the asthma subjects identified in ARIC were phenotypically distinct from the COPD subjects identified in ARIC. P-values are determined with Chi-Square or Fisher's exact test, where appropriate. Taken together, these results argue that our method of defining asthma by self-report combined with the exclusion of those with self-reported COPD successfully allowed us to isolate a cohort of asthmatics that was phenotypically distinct from subjects with COPD in ARIC.

Supplemental Table 10: Comparison of phenotypic variables between subjects with self-reported asthma or self-reported COPD in CARDIA. Key variables that distinguish patients with asthma from patients with COPD (allergic status, lung function, smoking status, and age) were compared in order to determine if the asthma subjects identified in CARDIA were phenotypically distinct from the COPD subjects identified in CARDIA. P-values are determined with Chi-Square or Fisher's exact test, where appropriate. Taken together, these results argue that our method of defining asthma by self-report combined with the exclusion of those with self-reported COPD successfully allowed us to isolate a cohort of asthmatics that was phenotypically distinct from subjects with COPD in CARDIA.

Supplemental Table 11: Comparison of phenotypic variables between subjects with self-reported asthma or self-reported COPD in JHS. Key variables that distinguish patients with asthma from patients with COPD (lung function, smoking status, and age) were compared in order to determine if the asthma subjects identified in JHS were phenotypically distinct from the COPD subjects identified in JHS. P-values are determined with Chi-Square or Fisher's exact test, where appropriate. Taken together, these results argue that our method of defining asthma by self-report combined with the

exclusion of those with self-reported COPD successfully allowed us to isolate a cohort of asthmatics that was phenotypically distinct from subjects with COPD in JHS.

Supplemental Table 12: Comparison of phenotypic variables between subjects with self-reported asthma or self-reported COPD in MESA. Key variables that distinguish patients with asthma from patients with COPD (allergic status, lung function, smoking status, and age) were compared in order to determine if the asthma subjects identified in MESA were phenotypically distinct from the COPD subjects identified in MESA. P-values are determined with Chi-Square or Fisher's exact test, where appropriate. Taken together, these results argue that our method of defining asthma by self-report combined with the exclusion of those with self-reported COPD successfully allowed us to isolate a cohort of asthmatics that was phenotypically distinct from subjects with COPD in MESA.

Supplemental Figure 1: Quantile-quantile (QQ) plots of the meta-analysis in the CARe African-American samples. Each black circle represents an observed statistic for all genotypes and imputed SNPs against the corresponding expected statistic. The grey area corresponds to the 90% confidence intervals calculated empirically using permutations.

Supplemental Figure 2: Quantile-quantile (QQ) plots of each individual CARe study used in the meta-analysis. Each black circle represents an observed statistic for all genotypes and imputed SNPs against the corresponding expected statistic. The area enclosed by the red lines corresponds to the 90% confidence intervals calculated empirically using permutations. λ_{1000} refers to λ for 1000 Genomes imputation.

Supplemental Figure 3: Fine mapping of association results at ORMDL3 locus. Asthma association results in Europeans ancestry individuals (A) and African Ancestry individuals plotted using LocusZoom against position on chromosome 5. The SNP name shown on the plot was the most significant SNP after meta-analysis. Estimated

recombination rates are plotted in cyan to reflect the local LD structure. The most significant SNP is colored purple. The SNPs surrounding the most significant SNP are color coded to reflect their LD with this SNP. Genes, the position of exons and the direction of transcription from the UCSC genome browser are noted. Hashmarks represent SNP positions available in the meta-analysis.

Supplemental References

Bild DE (2002) Multi-Ethnic Study of Atherosclerosis: Objectives and Design. American Journal of Epidemiology 156:871–881. doi: 10.1093/aje/kwf113

Lettre G, Palmer CD, Young T, et al. (2011) Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. PLoS Genet 7:e1001300. doi: 10.1371/journal.pgen.1001300.t004

Musunuru K, Lettre G, Young T, et al. (2010) Candidate Gene Association Resource (CARe): Design, Methods, and Proof of Concept. Circulation: Cardiovascular Genetics 3:267–275. doi: 10.1161/CIRCGENETICS.109.882696

Price AL, Patterson NJ, Plenge RM, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909. doi: 10.1038/ng1847

Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81:559–575. doi: 10.1086/519795