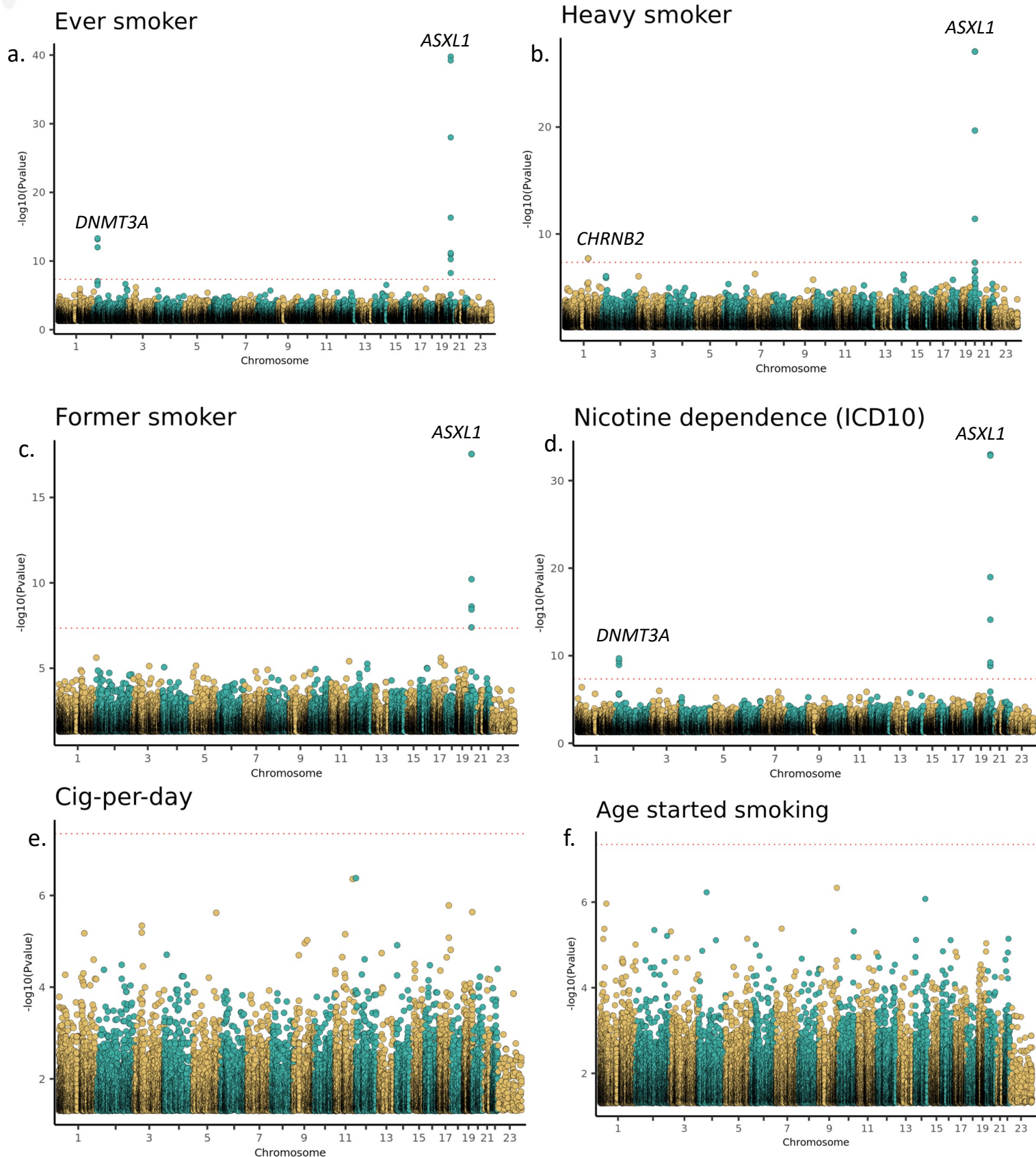
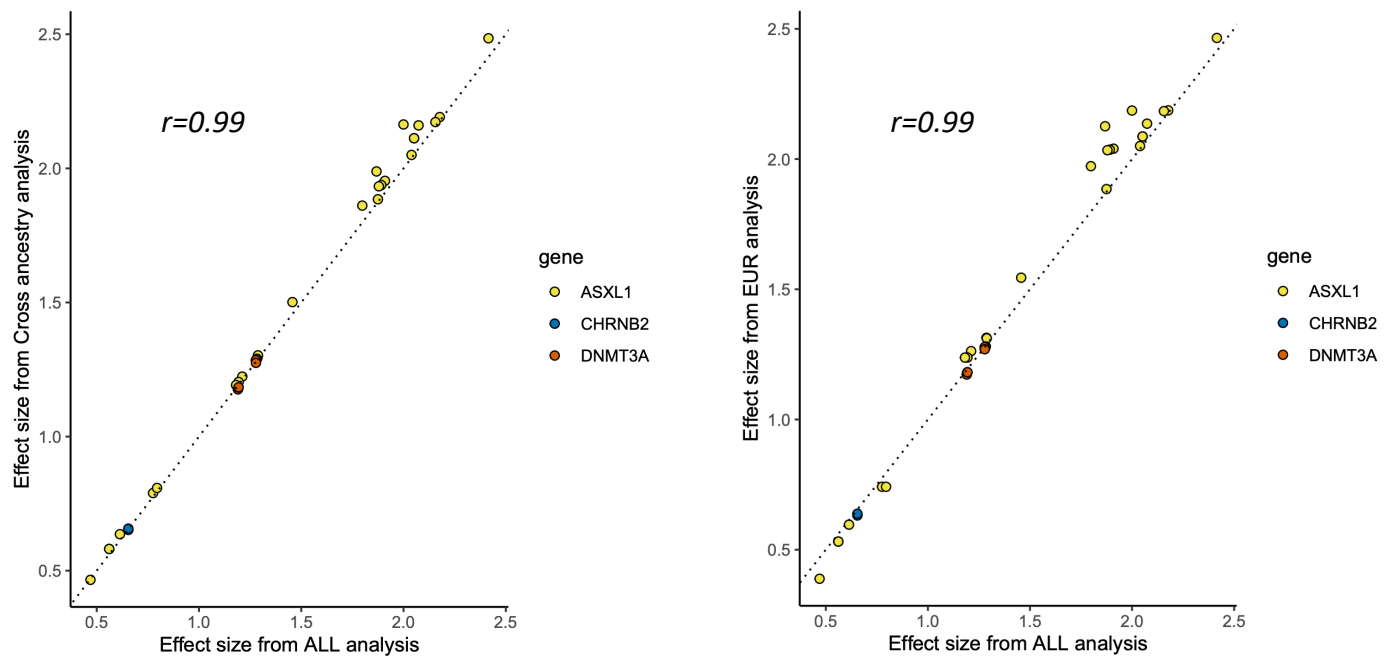


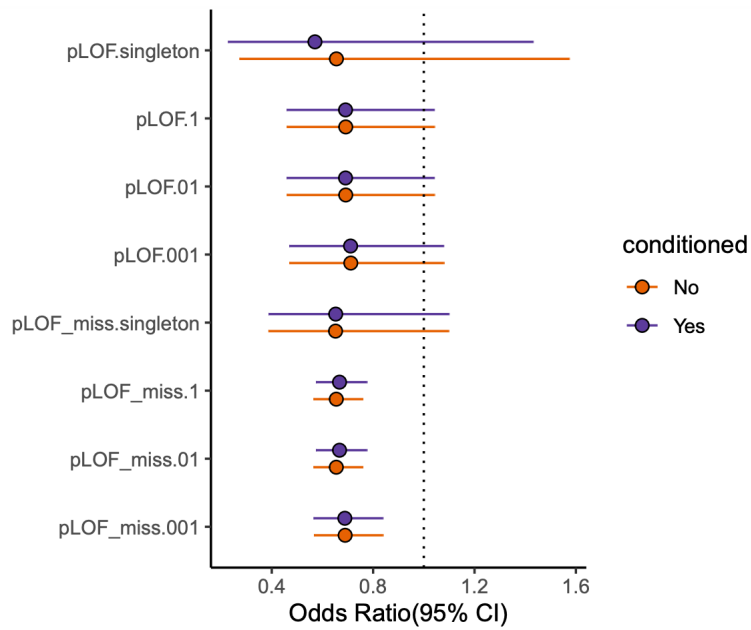
Supplementary Figure. 1. Overall study design



Supplementary Figure 2. ExWAS Manhattan plots of six smoking phenotypes. The plots display genome-wide genetic associations of both individual variants (loss of function and missense variants with MAF<0.01) and gene burden masks (pLOF only and pLOF plus likely deleterious missense variants at five MAF cut offs: <0.01, <0.001, <0.0001, <0.00001 and Singletons) with six primary smoking phenotypes (a-f). The exome-wide significance P value threshold was calculated by an applying a false detection rate (FDR) of 1% across the associations of all six smoking phenotypes (8,417,987 association tests in total), which corresponds to $P=4.5e-8$, marked with a dotted red line in each of the plots.

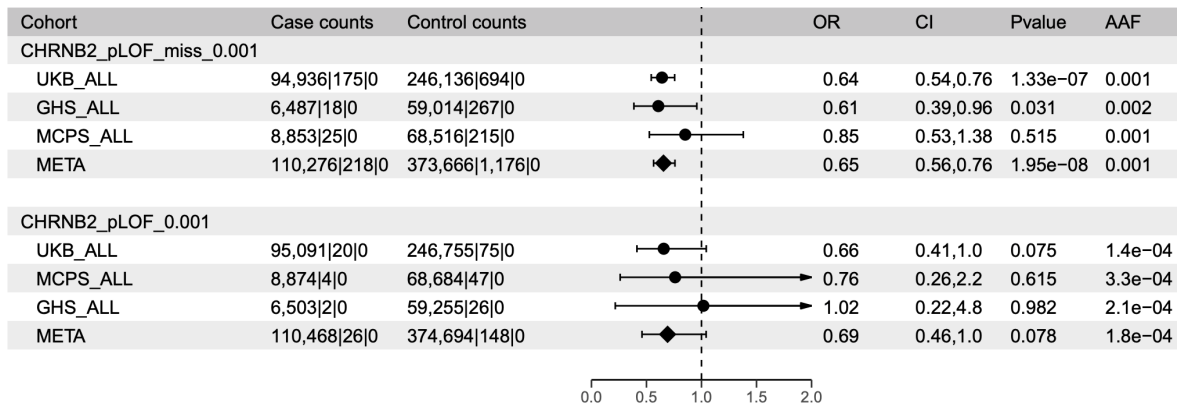


Supplementary Figure 3. Effect size (odds ratio) comparison of 35 significant ExWAS associations a. ALL (all ancestries pooled) meta-analysis vs cross ancestry meta-analysis (ancestry specific analysis followed by meta-analysis) and **b.** ALL meta-analysis vs EUR only meta-analysis. Pearson correlation (r) estimates are shown.

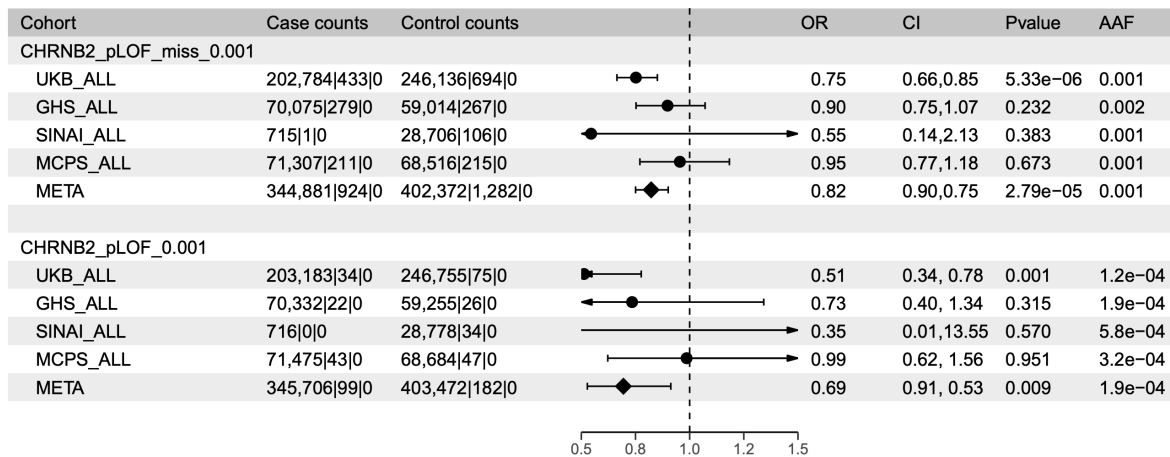


Supplementary Figure. 4. Analysis of *CHRN2* burden associations with heavy smoker conditioned on nearby common variants. *CHRN2* pLOF_miss_0.001 burden association with heavy smoker was computed before and after conditioning on nearby common variants recursively until no variants with 1 Mb on either side of the transcription start site of *CHRN2* has $P < 0.001$. The effect sizes of the burden associations remained the same after conditioning suggesting that the burden associations are independent of any nearby common variant GWAS signals.

a. ExWAS of heavy smoker

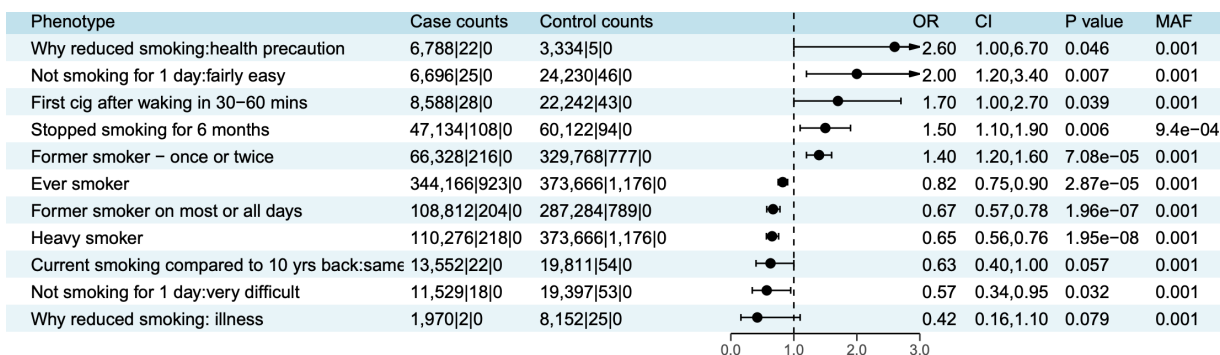


b. ExWAS of ever smoker

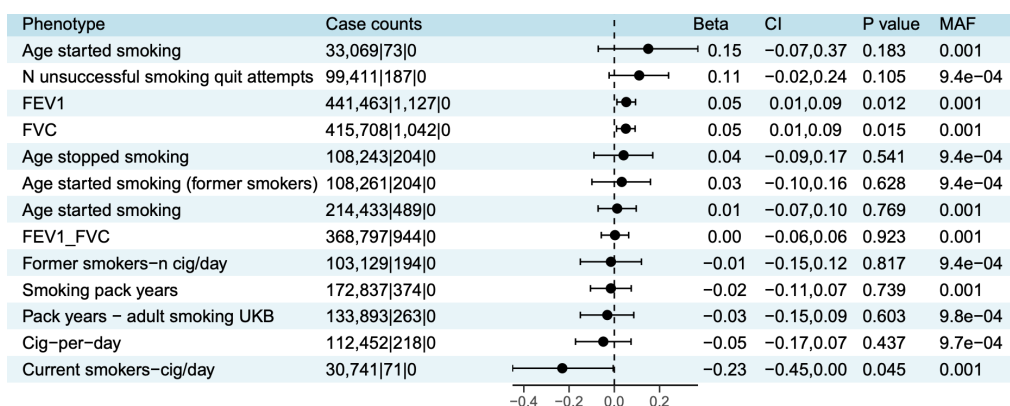


Supplementary Figure 5. Forest plots of associations of *CHRNA2* burden masks with a. heavy smoker and b. ever smoker. The genotype counts for reference allele homozygous, heterozygous and alternate allele homozygous are shown for cases and controls separately under the columns ‘Case counts’ and ‘Control counts’ respectively. UKB – UK Biobank; GHS – Geisinger Health System; MCPS – Mexico City Prospective Study; OR – Odds Ratio; CI – 95% confidence intervals; AAF – Alternate allele frequency;

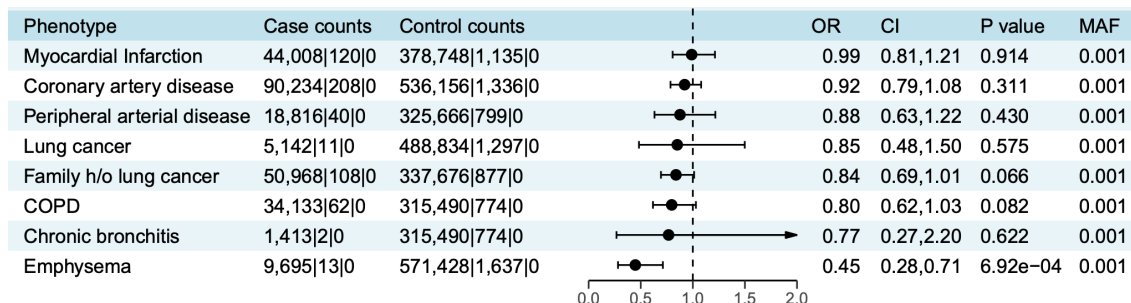
a.



b.

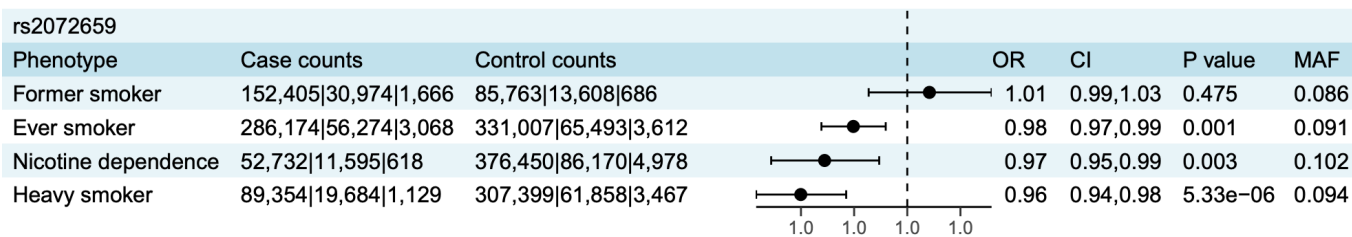


c.

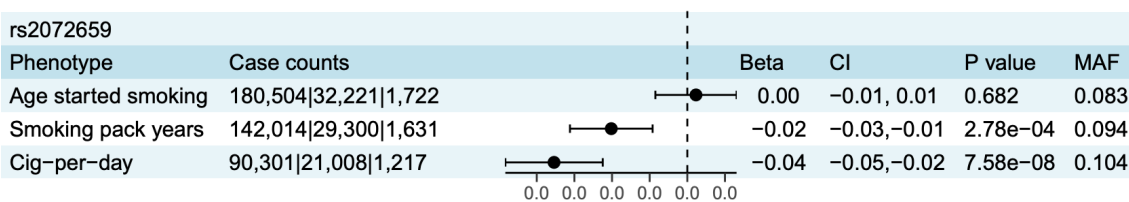


Supplementary Figure 6. Forest plots of PheWAS associations of *CHRN2* rare pLOF and missense burden with smoking
a. binary and **b.** quantitative phenotypes (six primary phenotypes plus lifestyle questionnaire derived secondary phenotypes in the UK Biobank; only associations with $P < 0.1$ are shown) and **c.** smoking related health conditions
 FEV1 – Forced expiratory volume in 1 sec; FVC – Forced vital capacity; FEV1_FVC – FEV1:FVC ratio; COPD – Chronic obstructive pulmonary disease; OR – Odds Ratio; CI – 95% Confidence Intervals; MAF – minor allele frequency

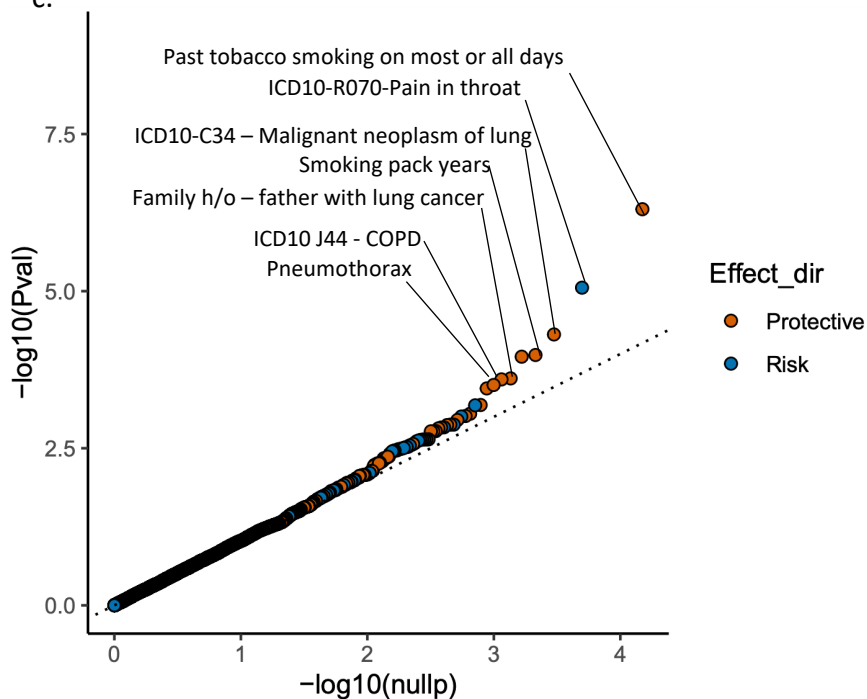
a.



b.

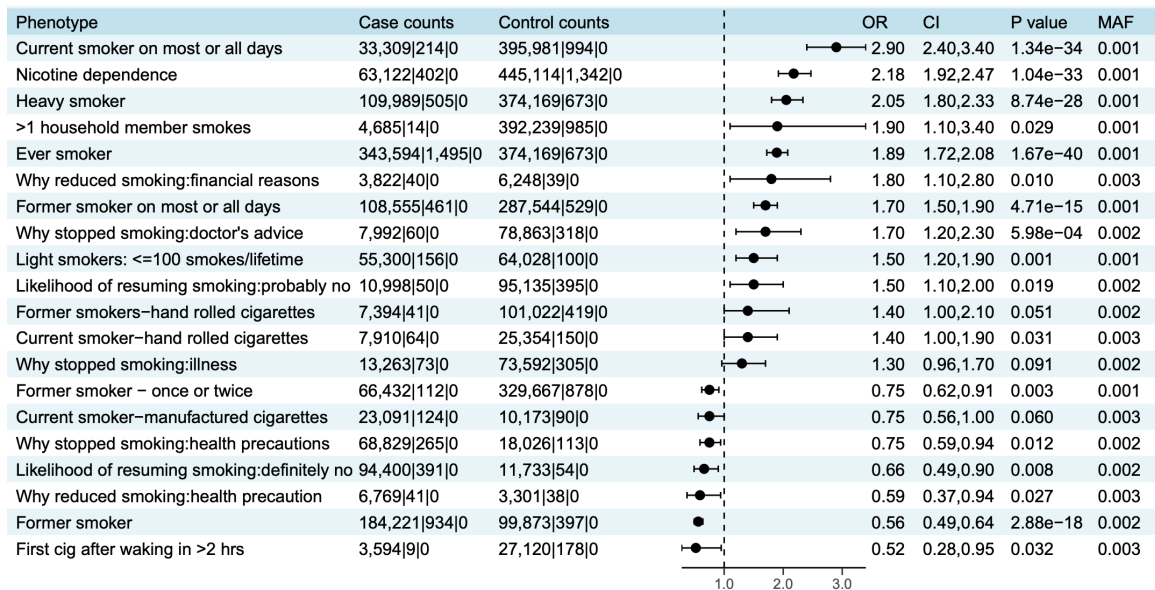


c.

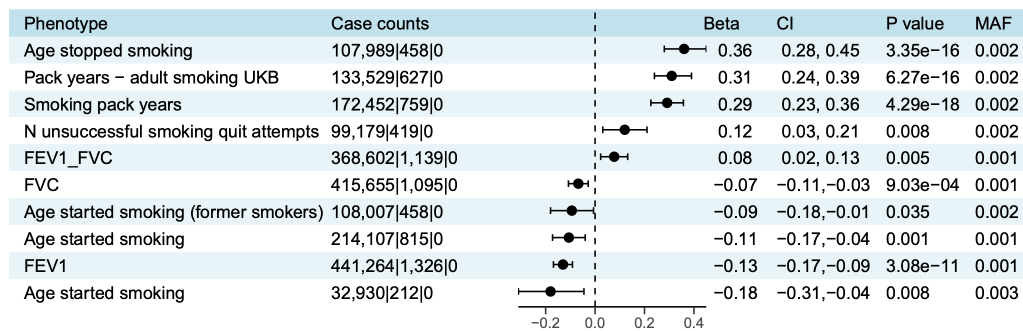


Supplementary Figure 7. Association of a *CHRN2* common 3' UTR variant with smoking. a. forest plots of associations with a. binary and b. quantitative smoking phenotypes. c. QQ plot of PheWAS of the common variant in UKB and GHS cohorts. The top 10 associations are labelled. Protective - OR <1 or beta <0 and risk - OR >1 or beta >0

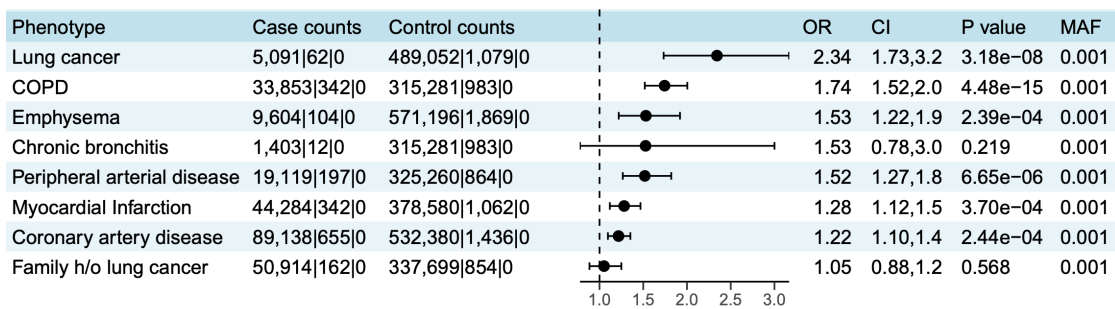
a.



b.

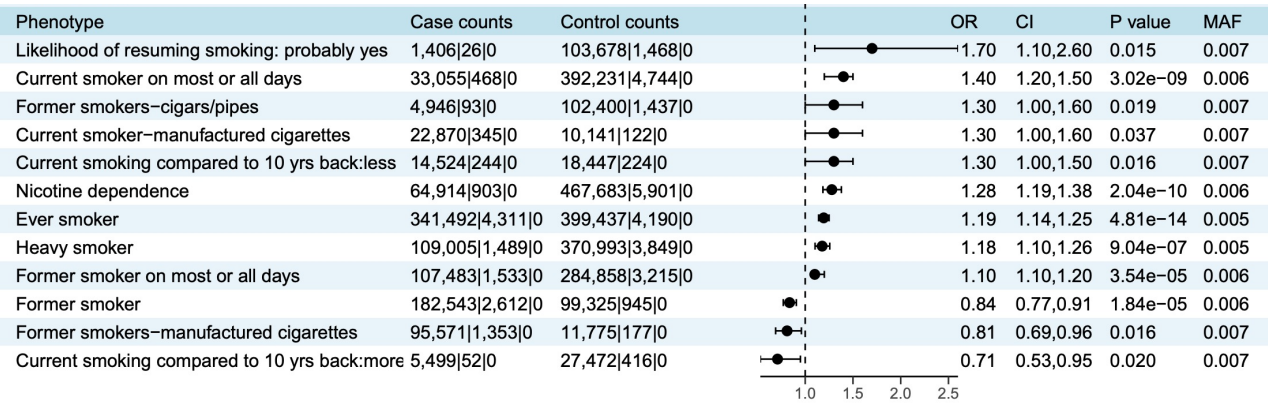


c.

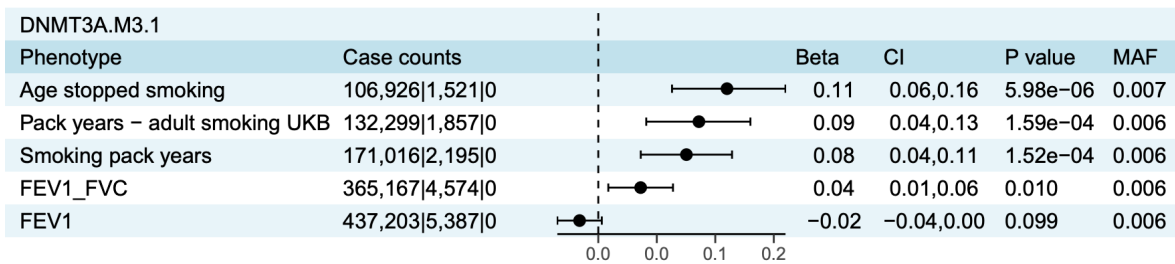


Supplementary Figure. 8. Forest plots of PheWAS associations of ASXL1 rare pLOF only burden with smoking a. binary and b. quantitative phenotypes (primary and secondary-questionnaire derived phenotypes in the UK Biobank; only associations with P<0.1 are shown) and c. smoking related health conditions

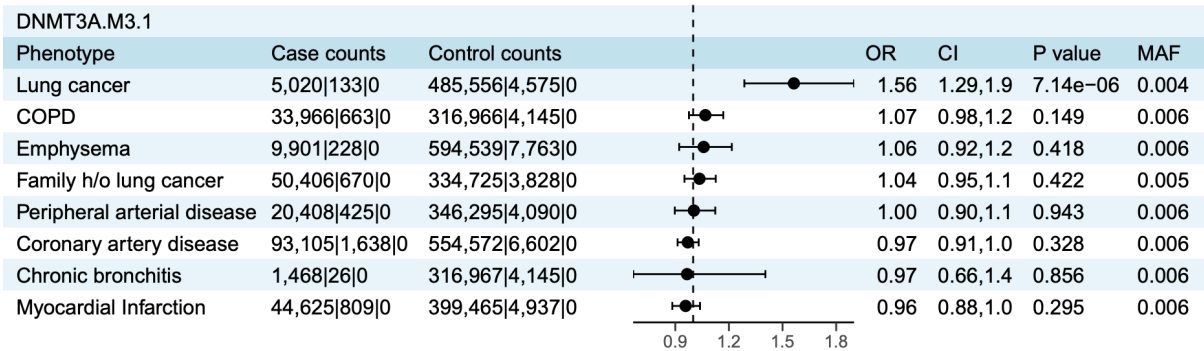
a.



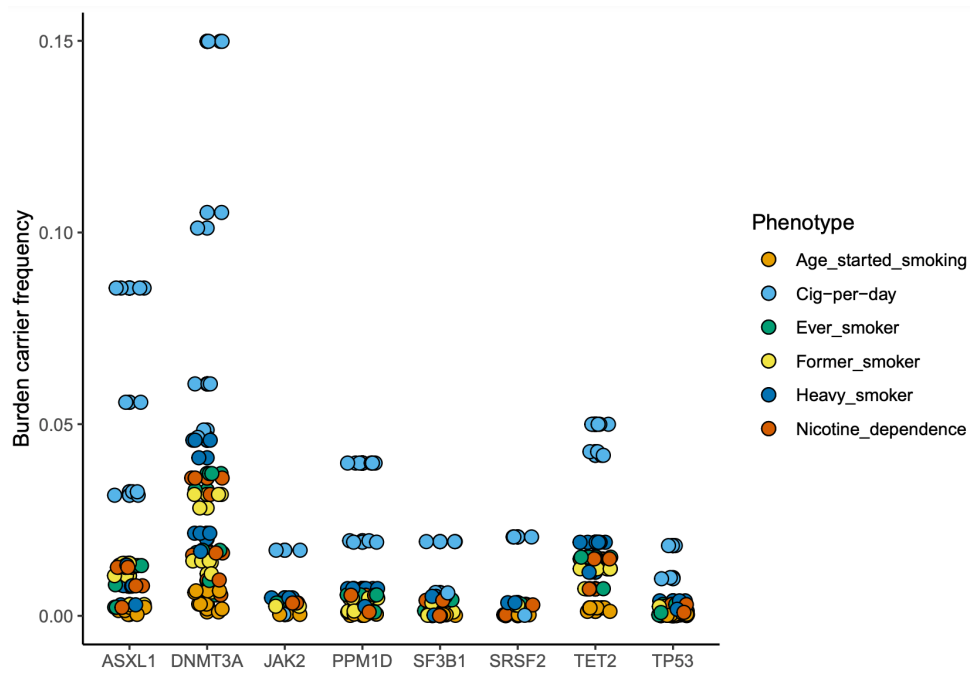
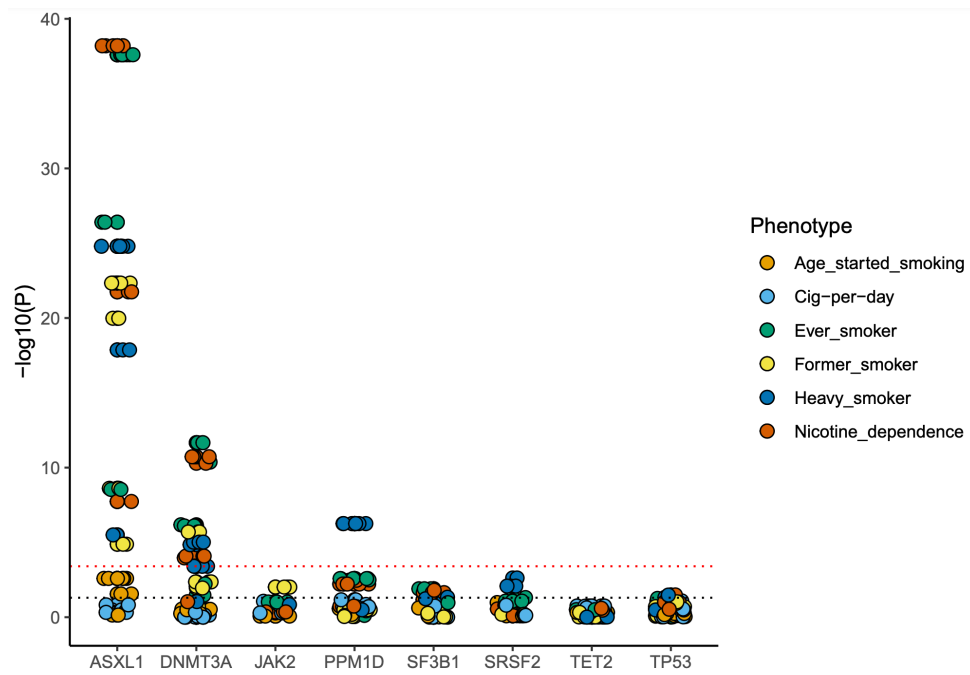
b.



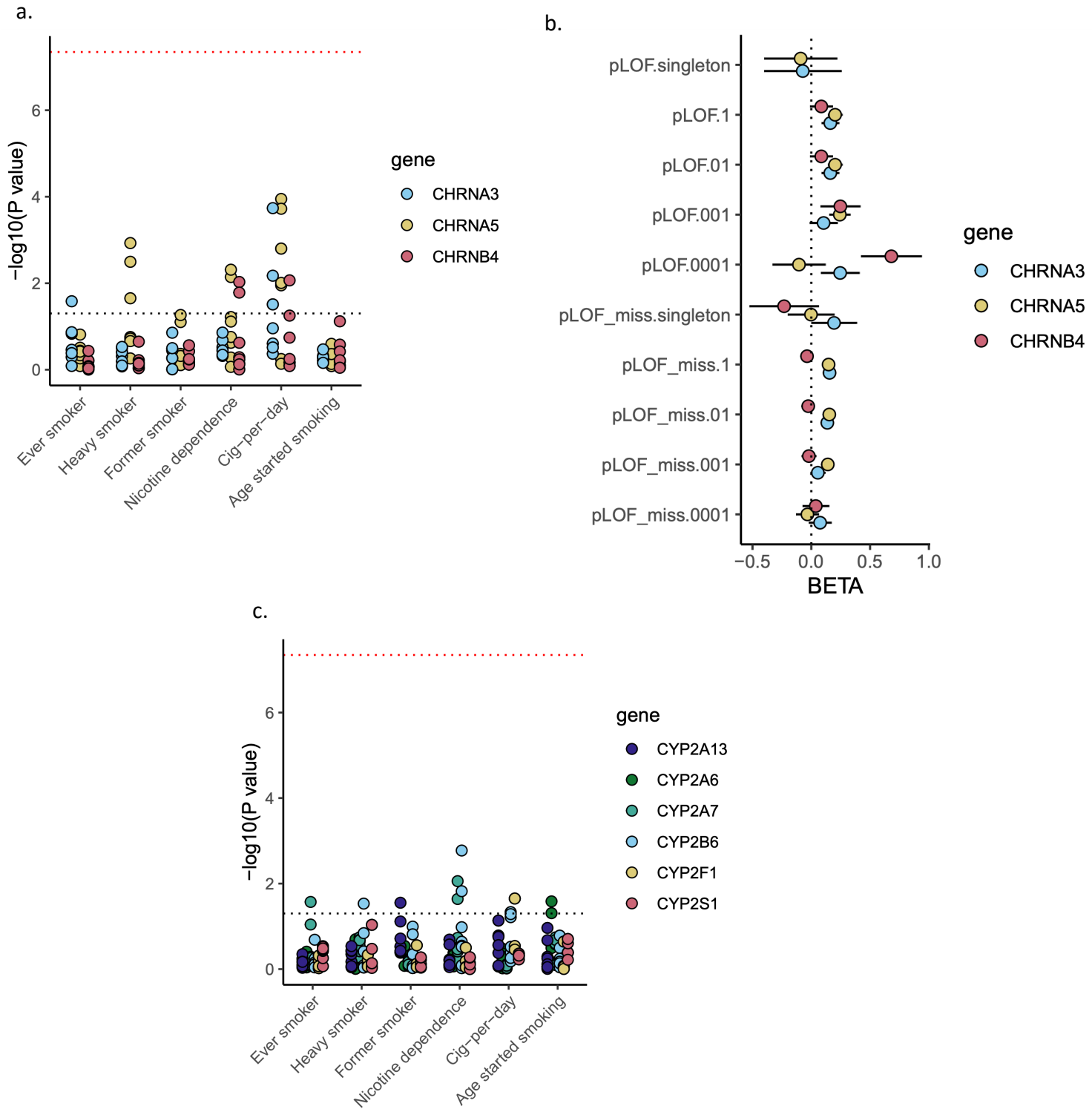
c.



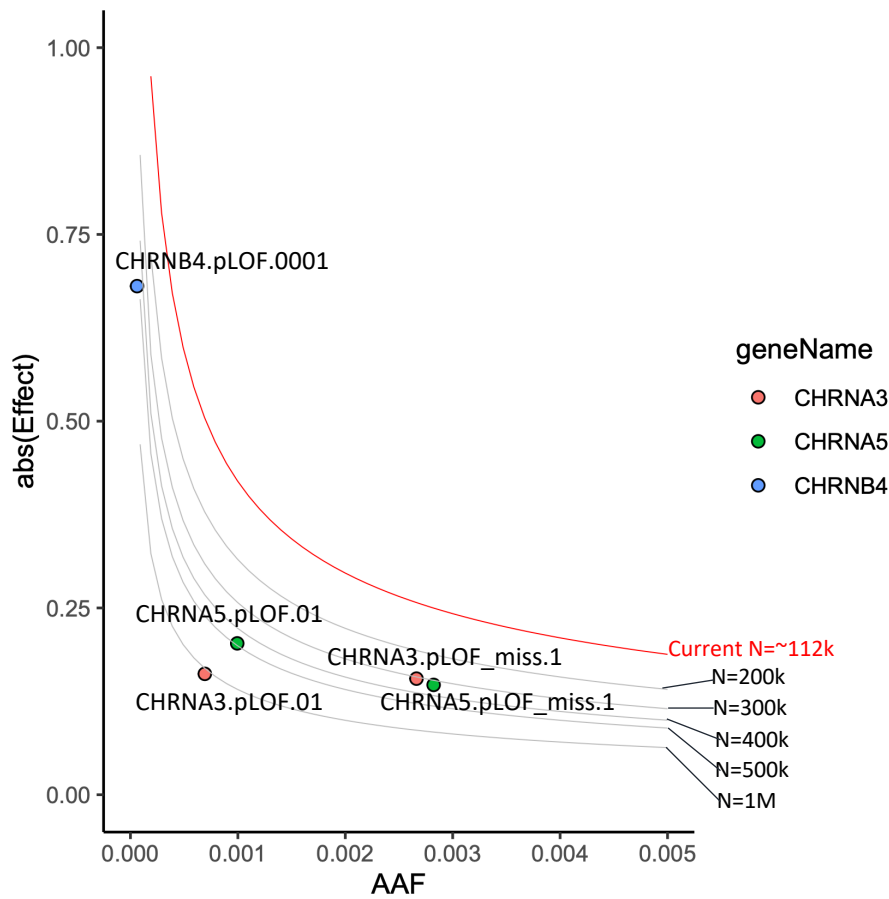
Supplementary Figure 9. Forest plots of PheWAS associations of DNMT3A rare pLOF plus missense burden with smoking
a. binary and **b.** quantitative phenotypes (primary and secondary-questionnaire derived phenotypes in the UK Biobank; only associations with $P < 0.1$ are shown) and **c.** smoking related health conditions



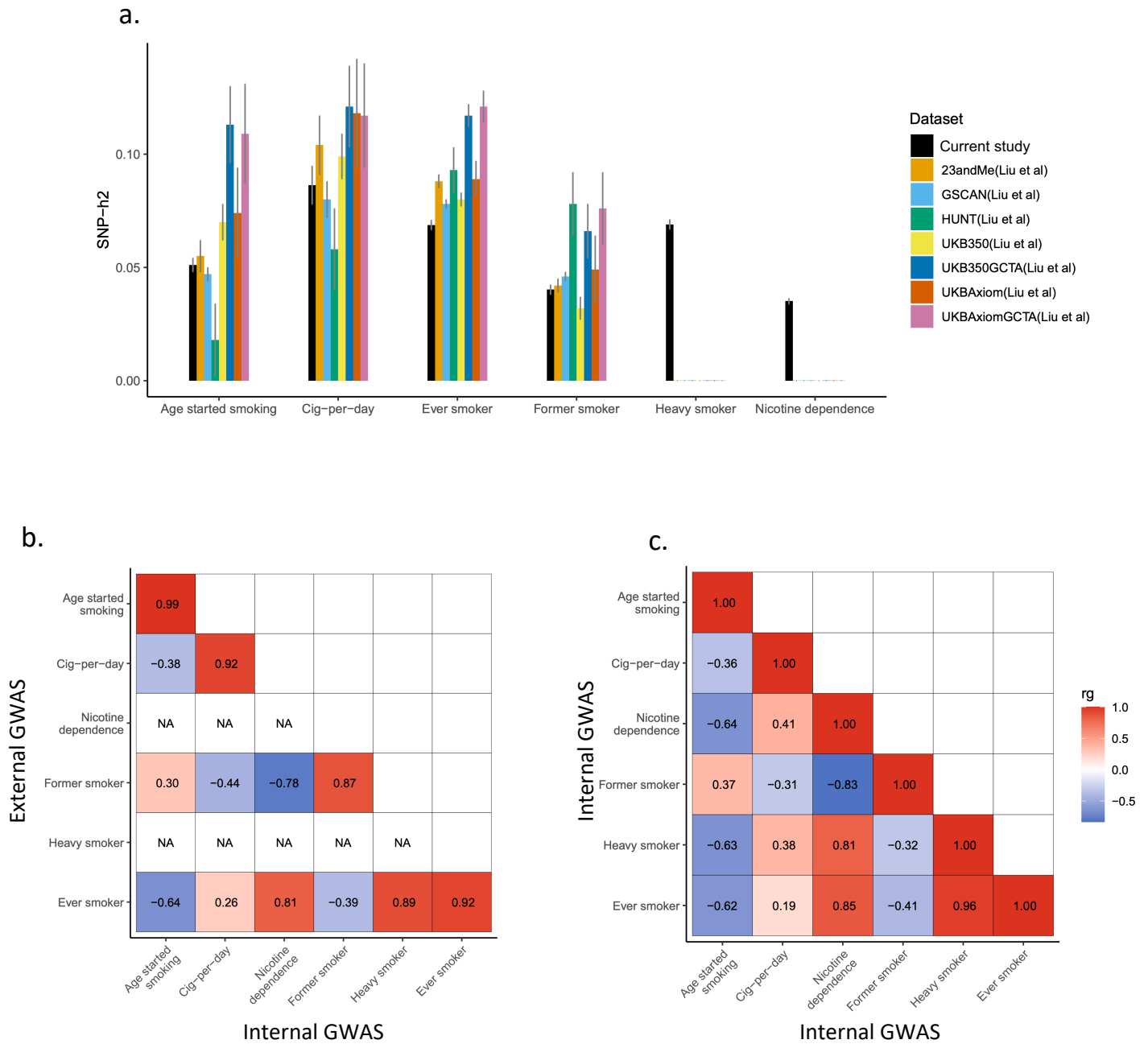
Supplementary Figure 10. Burden associations of eight most recurrent CHIP genes with smoking. The burden masks are created by collapsing only CHIP mutations (identified using a somatic mutation caller) under different definitions (Supplementary Table 2) and testing their associations with the six smoking phenotypes. The P values of the associations are plotted in **a.** in $-\log_{10}$ scale. The red dotted line corresponds to a $P=0.0003$ at FDR 1% threshold applied across all the eight genes. The black dotted line corresponds to $P=0.05$. The burden allele frequencies are plotted in **b.** to illustrate that some genes such as TET2, despite harboring CHIP mutations in high frequency, did not show significant associations with smoking.



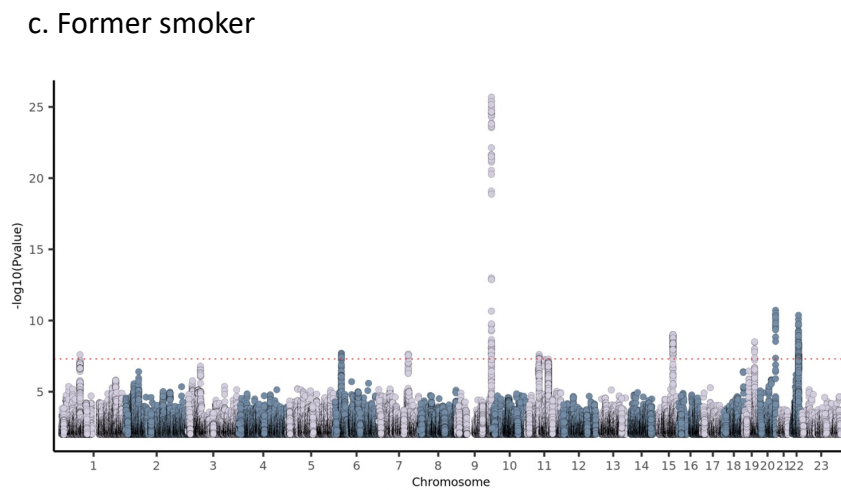
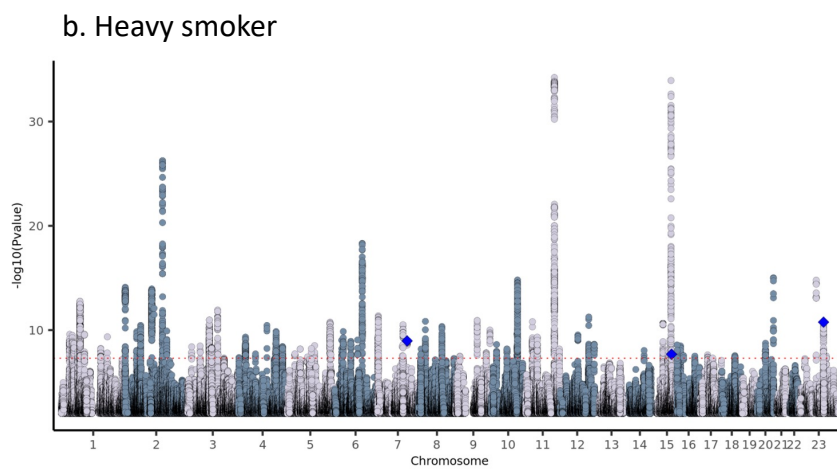
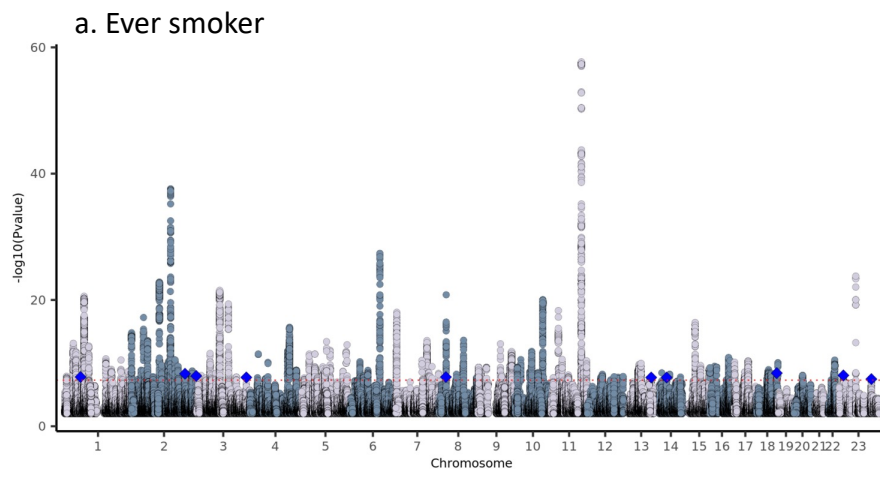
Supplementary Figure 11. Rare variant associations at the classic GWAS loci. a. Associations of 10 burden masks (Supplementary Table 2) of *CHRNA5*, *CHRNA3* and *CHRNA4* (at smoking GWAS locus 15q25.1) with six smoking phenotypes are plotted. The red dotted line corresponds to exome-wide significant threshold P value=4.5e-8. The black dotted line corresponds to P=0.05. **b.** the effect sizes (Beta values) and standard errors corresponding to the data points in fig a. are plotted. **c.** Associations of 10 burden masks of CYP450 family of genes at smoking GWAS locus 19q13.2 are plotted.



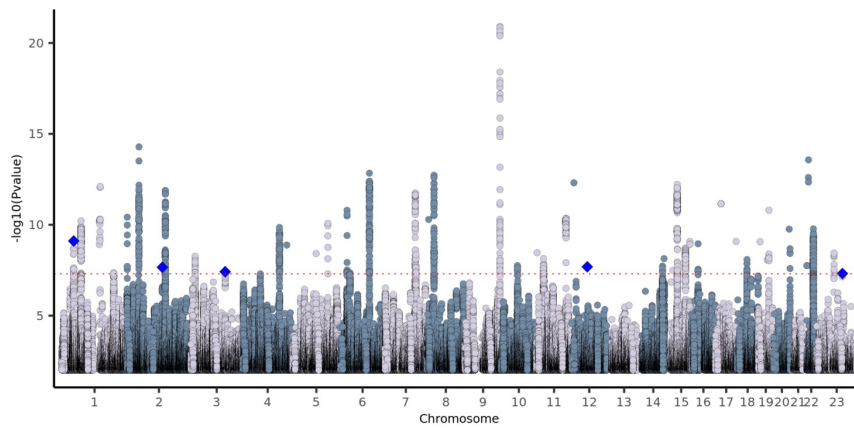
Supplementary Figure 12. Power calculations for rare variant discovery at the CHRNA5 GWAS locus. Assuming an 80% power and P value of $5e-8$, detectable effect sizes at various minor allele frequency values were calculated for the current sample size of cig-per-day (the smoking trait most associated with CHRNA5 locus) as well for a series of sample sizes up to 1 million. The observed effect sizes for pLOF only burden and pLOF and missense burden associations of CHRNA5, CHRNA3 and CHRNB4 are plotted; all the points lay below the red line, which marks the detection limit of our current sample size, suggesting that we are underpowered. Based on the intersections of the grey lines with the points marking the observed effect sizes, we can approximately guess what sample size will be required to detect these burden signals at P value $5e-8$.



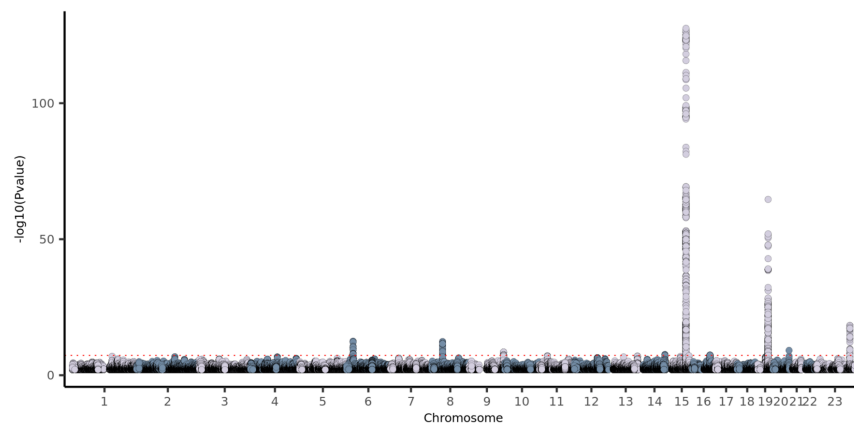
Supplementary Figure 13. SNP-h2 and genetic correlations. a. SNP-h2 of the six smoking phenotypes estimated using LD score regression are plotted along with standard errors. The SNP-h2 estimates reported by Liu et al 2019 are also plotted along side for comparisons. Note, Liu et al did not study heavy smoker and nicotine dependence phenotypes. b. Genetic correlations between our six smoking phenotypes and the four smoking phenotypes from GSCAN consortium (Liu et al 2019) estimated using LD score regression are shown as heat maps. The r_g estimates are displayed over the plots. Nicotine dependence and heavy smoker phenotypes are not studied by Liu et al, hence, shown as NA. c. Genetic correlations between the smoking phenotypes within our cohort are displayed. Internal GWAS – Cohorts involved in the current study
External GWAS – GSCAN cohorts (excluding UKB)



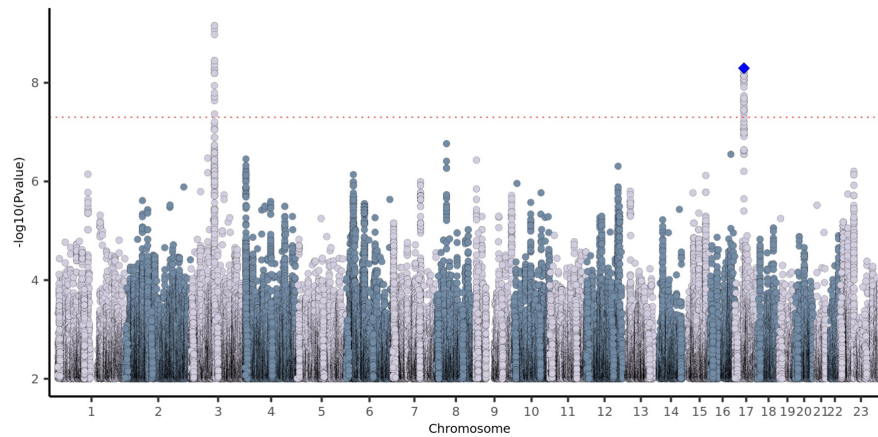
d. Nicotine dependence



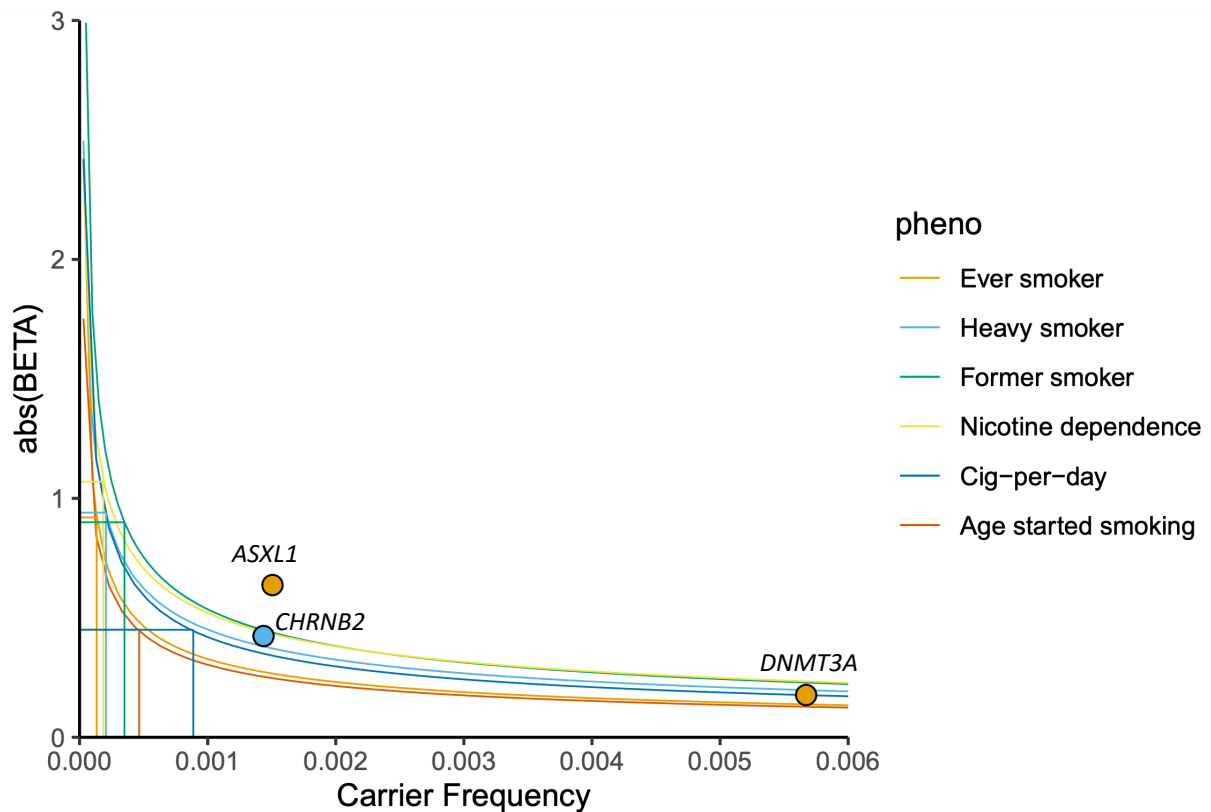
e. Cig-per-day



f. Age started smoking



Supplementary Figure. 14 Manhattan plots of cross-ancestry GWAS meta-analyses of the six smoking phenotypes (a-f). The novel GWAS loci are marked with blue diamonds in the Manhattan plots. The red dotted line corresponds to P value of $5e-8$.



Supplementary Figure 15. Power calculations for gene discovery using the current sample size. Assuming an 80% power, P value threshold of $4e-8$ (exome-wide significant threshold of the current study based on FDR 1%), effect sizes (i.e., beta values) were computed for a range of minor allele frequencies (combined allele frequency in case of burden masks) for a given sample size (varies across phenotypes). The computed effect sizes (absolute values of beta estimates) are plotted against minor allele frequencies (carrier frequency) for six smoking phenotypes. The carrier frequency corresponding to 100 carriers, calculated for each of the phenotype based on the corresponding sample size, in the X axis and the corresponding effect size in the Y axis are marked with straight lines. The top association of the three genes identified as exome-wide significant are plotted with the color corresponding to the associated phenotype. Based on these power curves, we had 80% power to detect any variant or burden associations with ever smoker, heavy smoker and former smoker with odds ratio ~ 2.5 or higher (0.4 or lower) when there are at least 100 carriers. And we had 80% power to detect any variant or burden associations with cig-per-day and age started smoking with beta 0.45 (equivalent to 4.7 extra cigarettes for cig-per-day and 1.9 earlier age for age started smoking) when there are at least 100 carriers. These calculations assume that there is no heterogeneity in the effect sizes across the cohorts, which is never the case for complex traits such as smoking. Hence, these estimates should be considered arbitrary. Importantly, the effect sizes for protective associations with binary phenotypes are likely overestimated due to imbalance in the case-control ratios.