

Choquet et al. “Multi-ancestry genome-wide meta-analysis identifies novel basal cell carcinoma loci and shared genetic effects with squamous cell carcinoma”

Supplementary Figures

Supplementary Figure 1. QQ plot and genomic inflation factor (λ) observed for the European ancestry GWA meta-analysis of BCC.

Supplementary Figure 2. Manhattan plot of the European ancestry GWA meta-analysis of BCC.

Supplementary Figure 3. Locus Zoom plots of novel regions identified in the European ancestry GWA meta-analysis of BCC.

Supplementary Figure 4. Forest plots of new BCC loci identified in the European ancestry GWA meta-analysis.

Supplementary Figure 5. QQ plot and genomic inflation factor (λ) observed for the Hispanic/Latino ancestry GWA meta-analysis of BCC.

Supplementary Figure 6. Manhattan plot of the Hispanic/Latino ancestry GWA meta-analysis of BCC.

Supplementary Figure 7. Locus Zoom plots of regions identified in the Hispanic/Latino GWA meta-analysis of BCC or showing association with BCC at Bonferroni level of significance.

Supplementary Figure 8. Receiver operating characteristic curves for BCC predictions in the GERA Hispanic/Latino sample, using polygenic risk scores derived from the European ancestry GWA meta-analysis of BCC.

Supplementary Figure 9. QQ plot and genomic inflation factor (λ) observed for the multi-ancestry GWA meta-analysis of BCC.

Supplementary Figure 10. Locus Zoom plots of additional novel regions identified in the multi-ancestry GWA meta-analysis of BCC.

Supplementary Figure 11. Correlation of effect sizes across populations (European ancestry vs. Hispanic/Latino) for the lead 116 BCC-associated lead SNPs identified in the European ancestry GWA meta-analysis of BCC.

Supplementary Figure 12. QQ plot and genomic inflation factor (λ) observed for the European ancestry GWA meta-analysis of SCC.

Supplementary Figure 13. Manhattan plot of the European ancestry GWA meta-analysis of SCC.

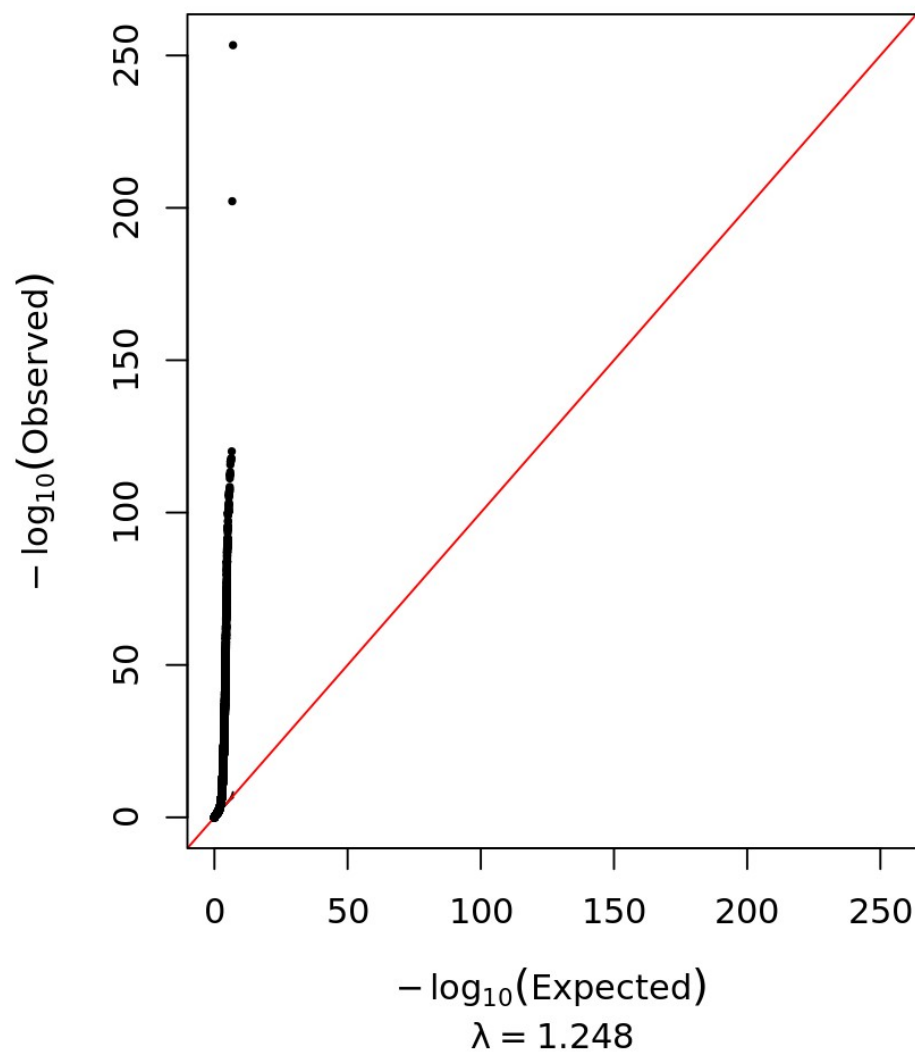
Supplementary Figure 14. Genome-wide genetic correlations between BCC and SCC across cohorts.

Supplementary Notes

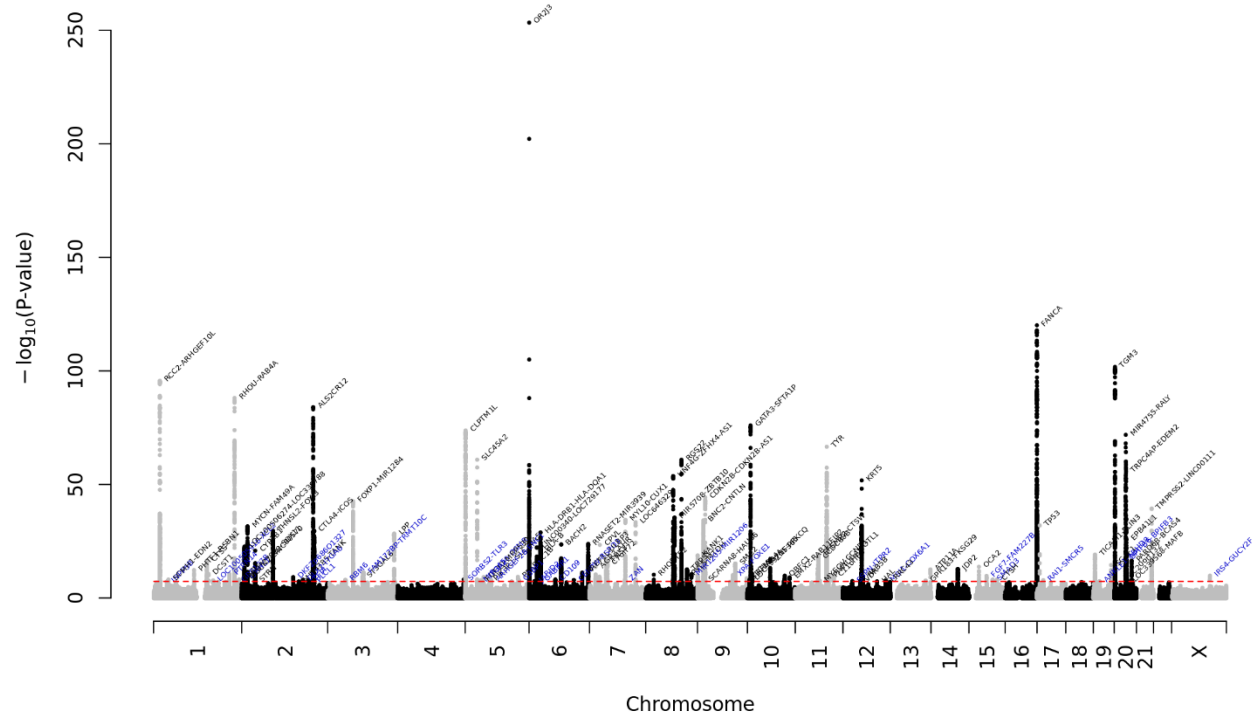
Supplementary Note 1: Details for UKB on phenotyping

Supplementary Note 2: Details for each cohort on genotyping, imputation, and quality control

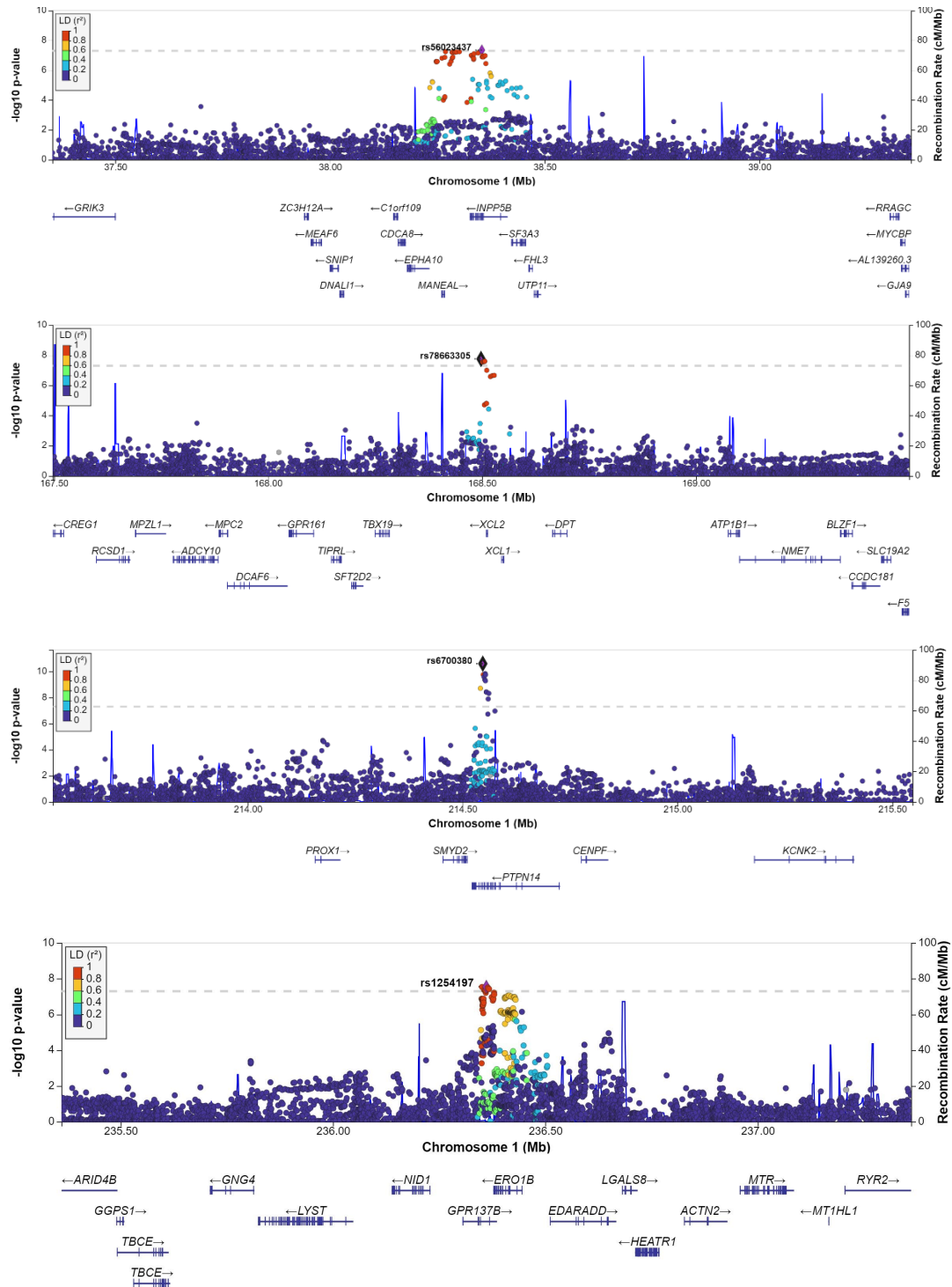
Supplementary Figure 1. QQ plot and genomic inflation factor (λ) observed for the European ancestry GWA meta-analysis of BCC

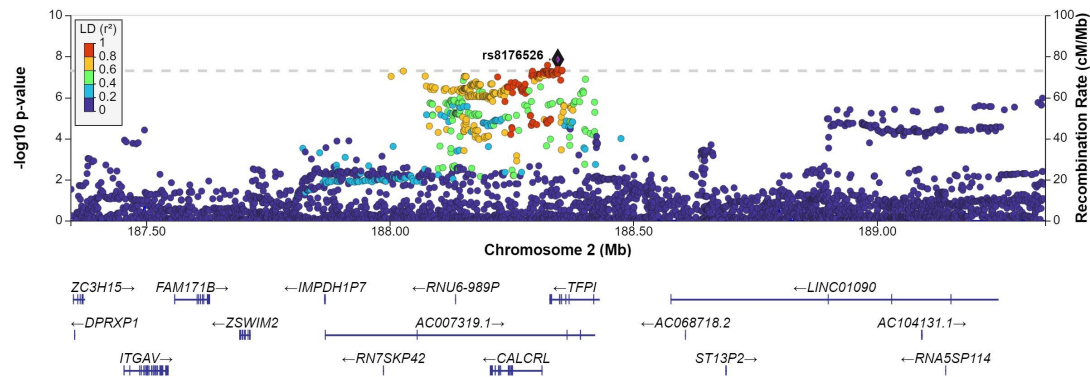
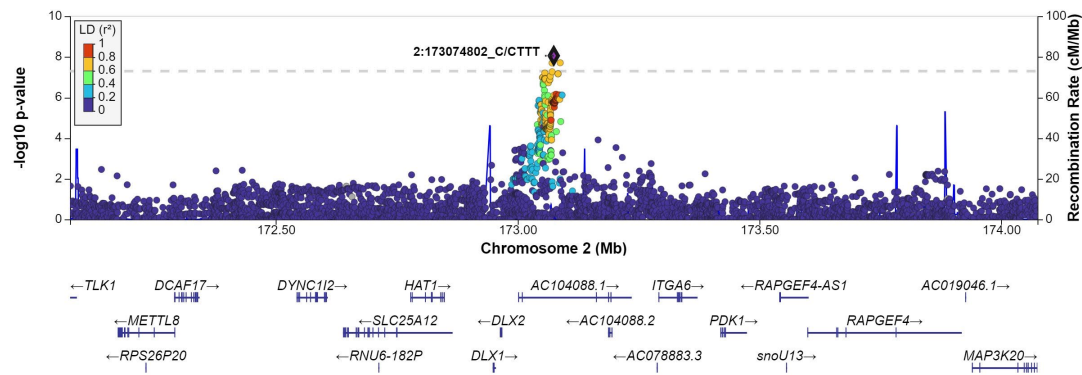
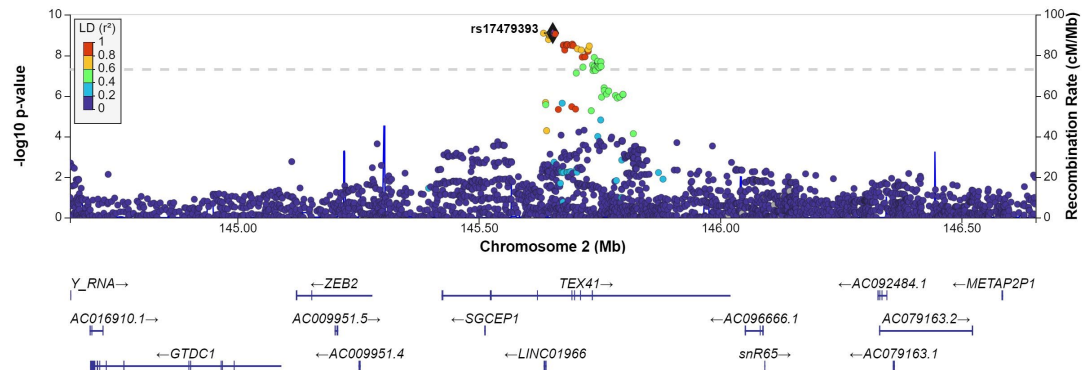
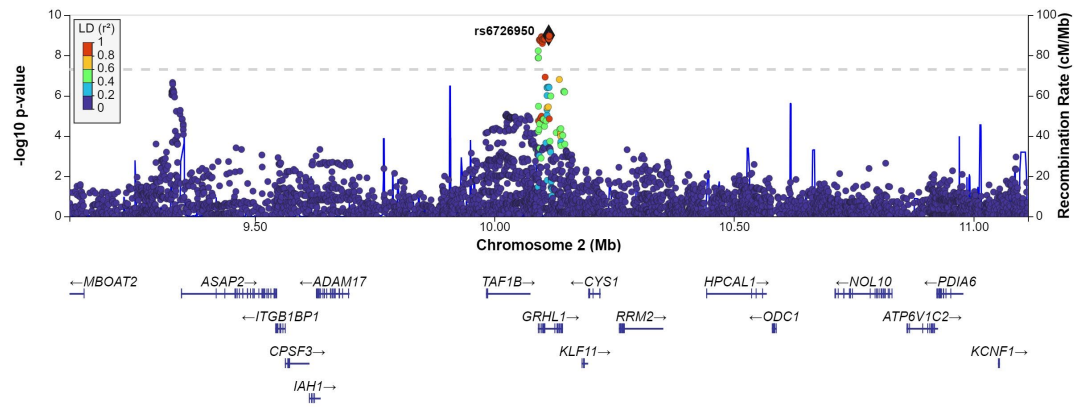


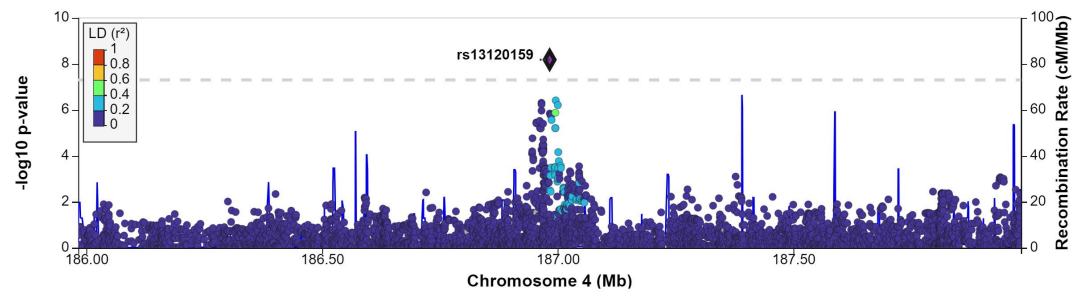
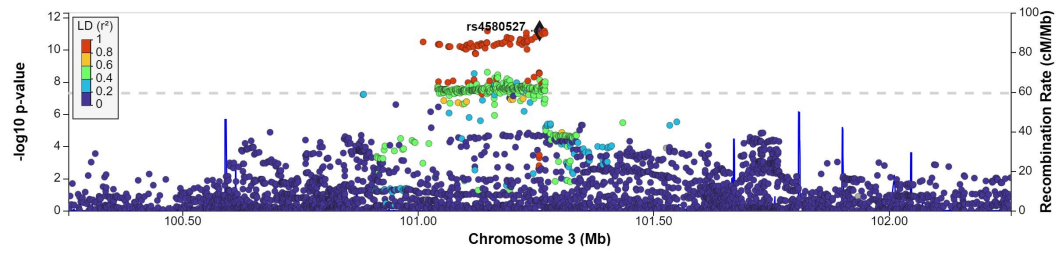
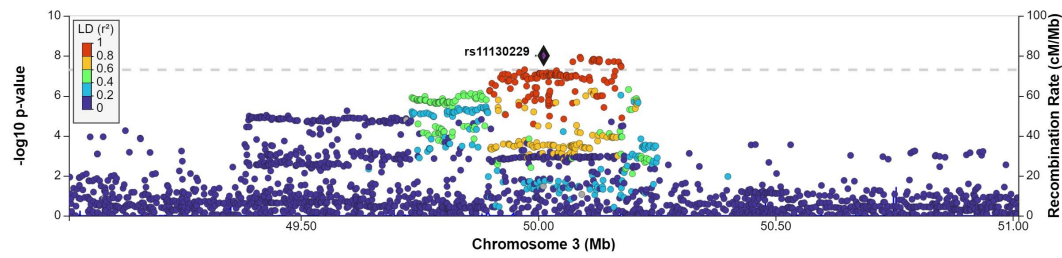
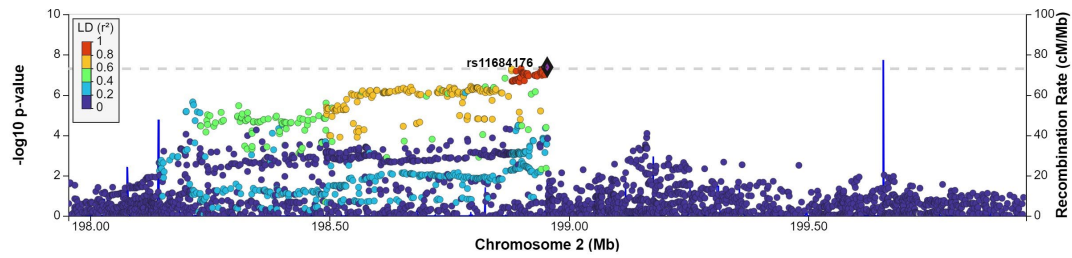
Supplementary Figure 2. Manhattan plot of the European ancestry GWA meta-analysis of BCC. The y-axis represents the $-\log_{10}(P\text{-value})$; all P -values derived from logistic regression model are two-sided. The red dotted line represents the threshold of $P=5 \times 10^{-8}$ which is the commonly accepted threshold of adjustments for multiple comparisons in GWAS. Locus names in blue are for the novel loci and the ones in dark are for the previously reported ones.

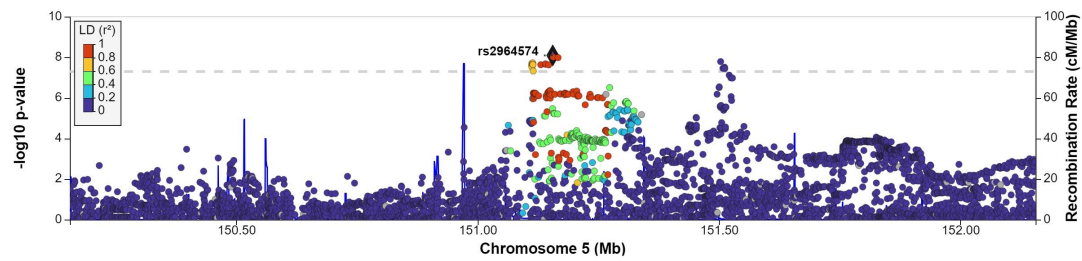
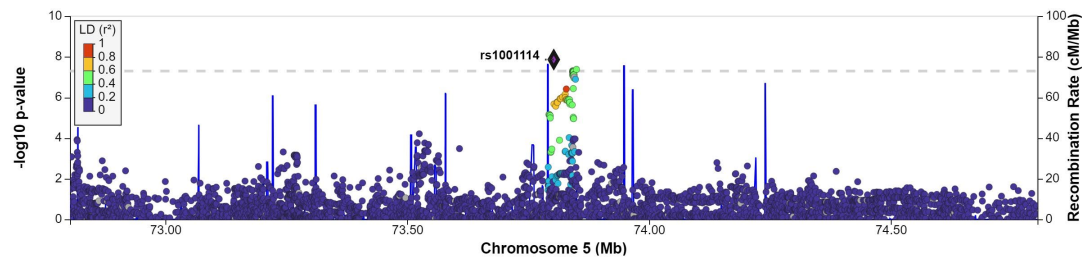
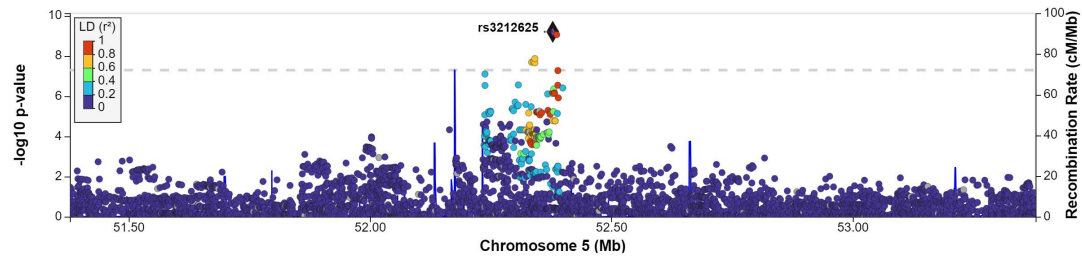
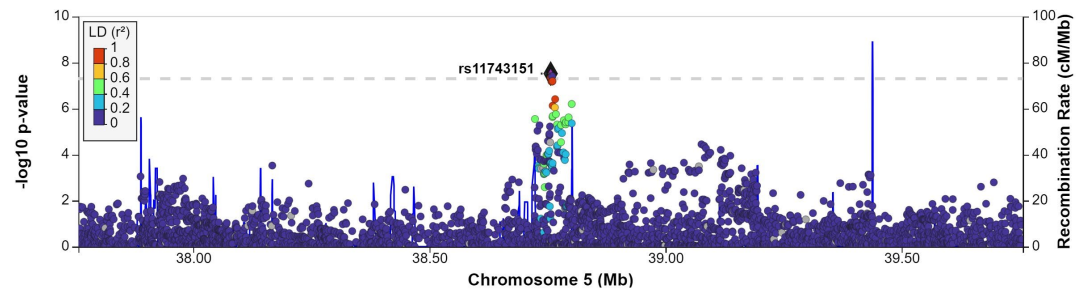


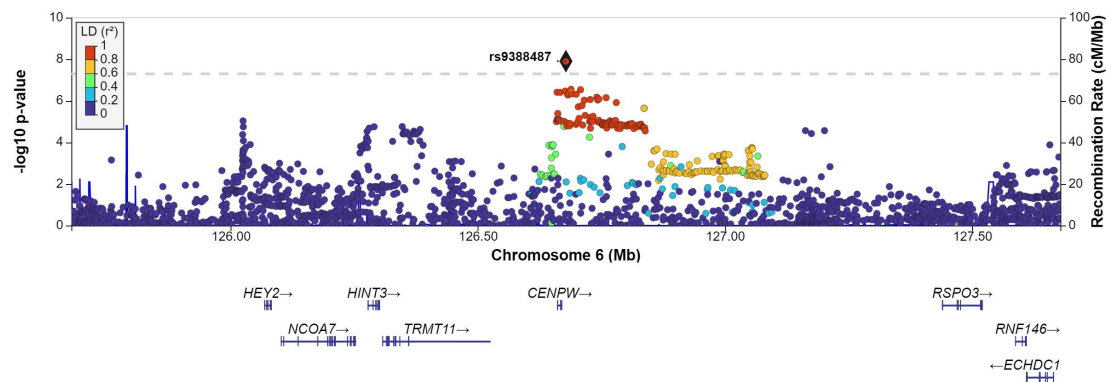
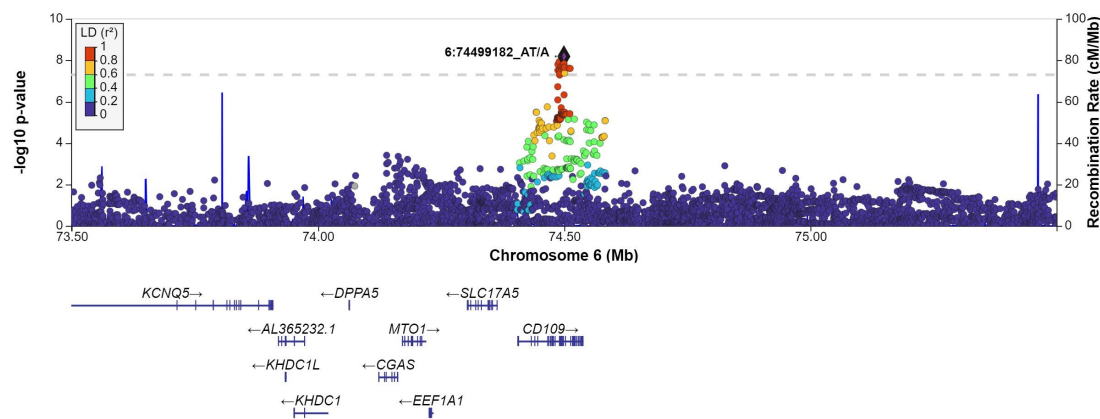
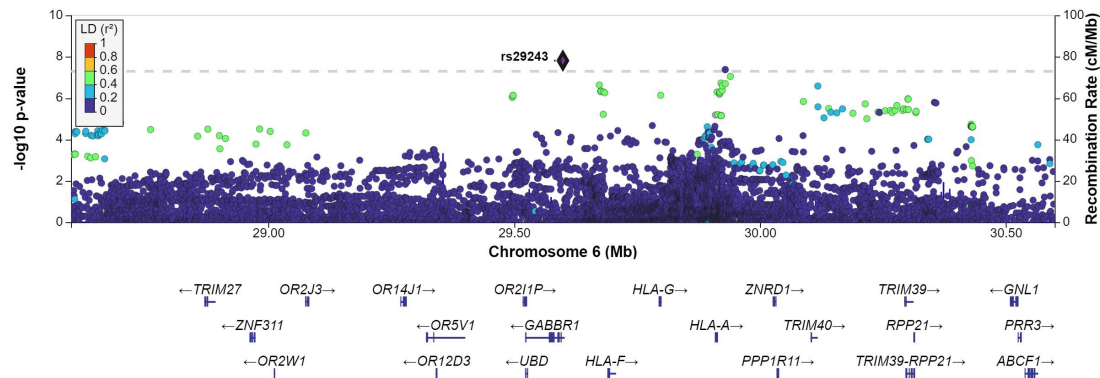
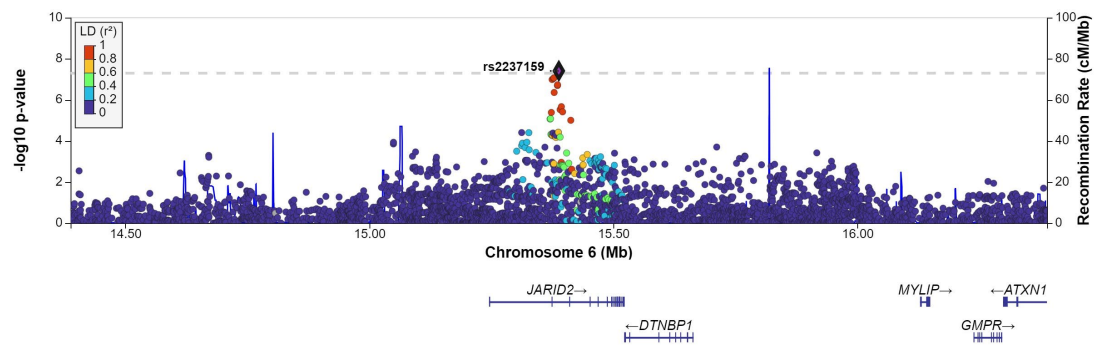
Supplementary Figure 3. Locus Zoom plots of novel regions identified in the European ancestry GWA meta-analysis of BCC.

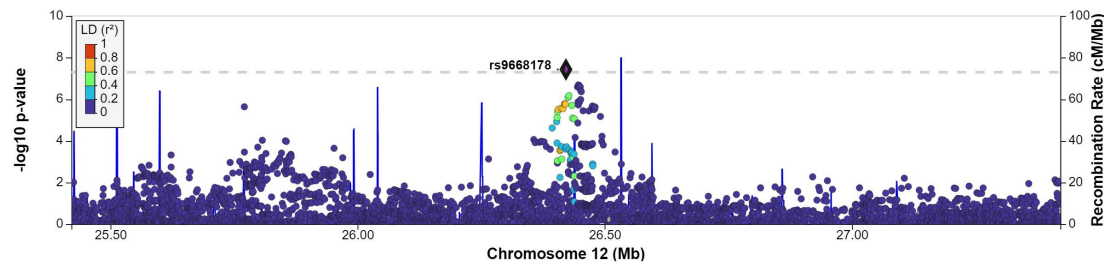
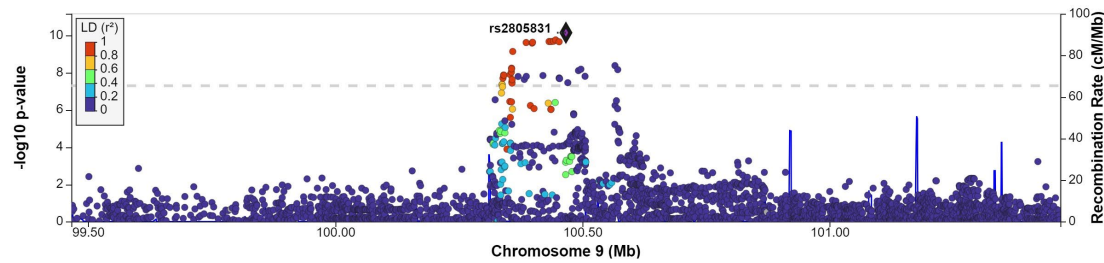
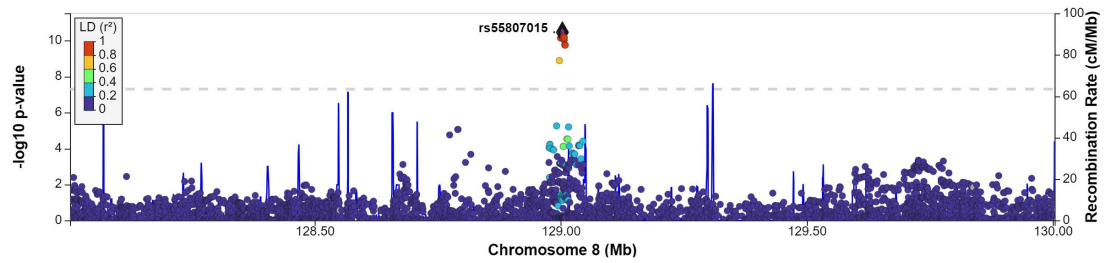
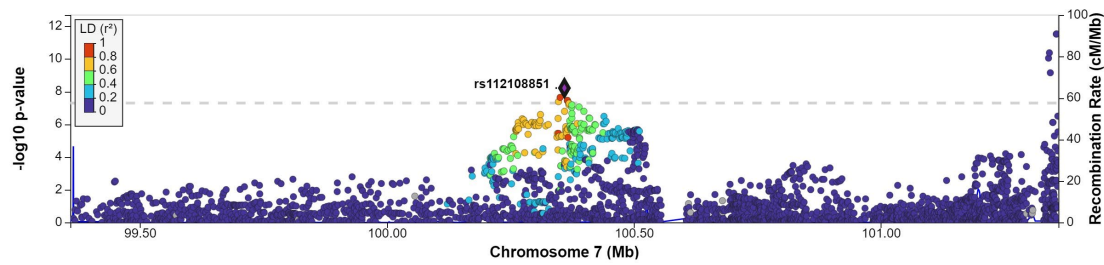


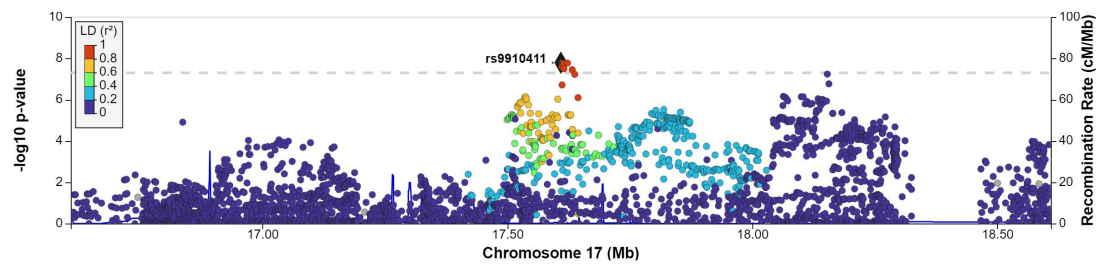
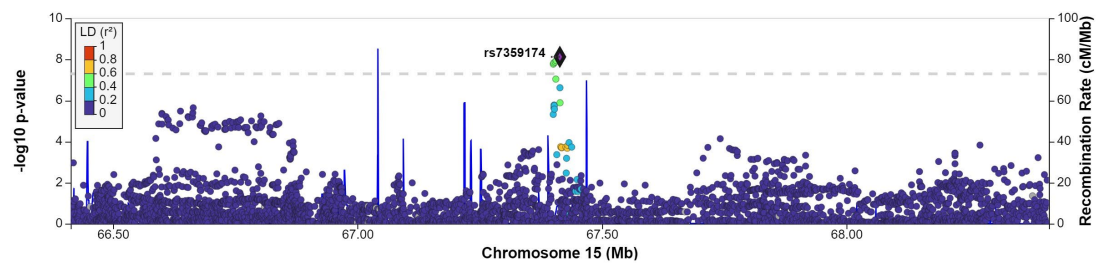
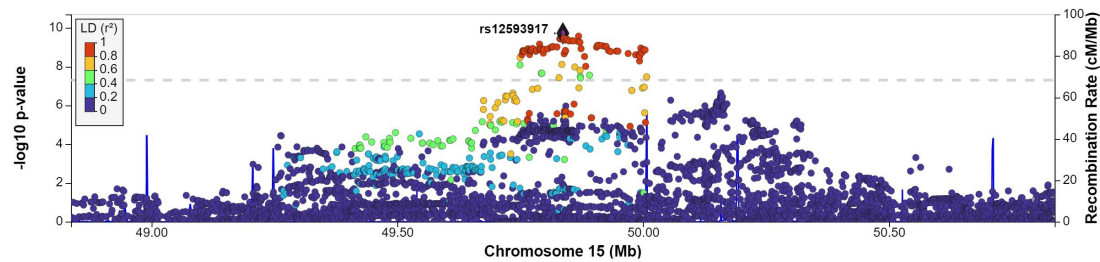
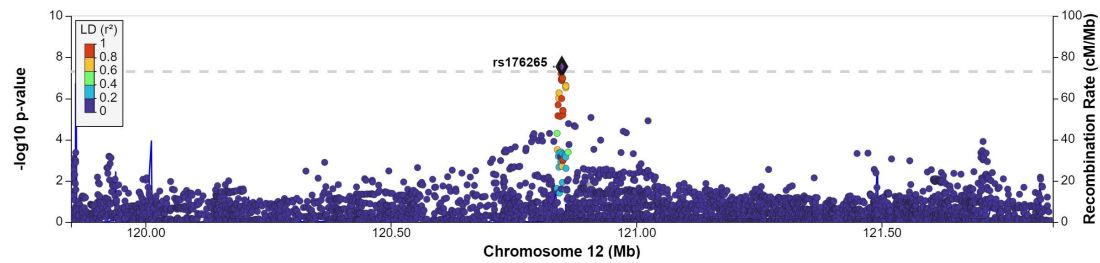


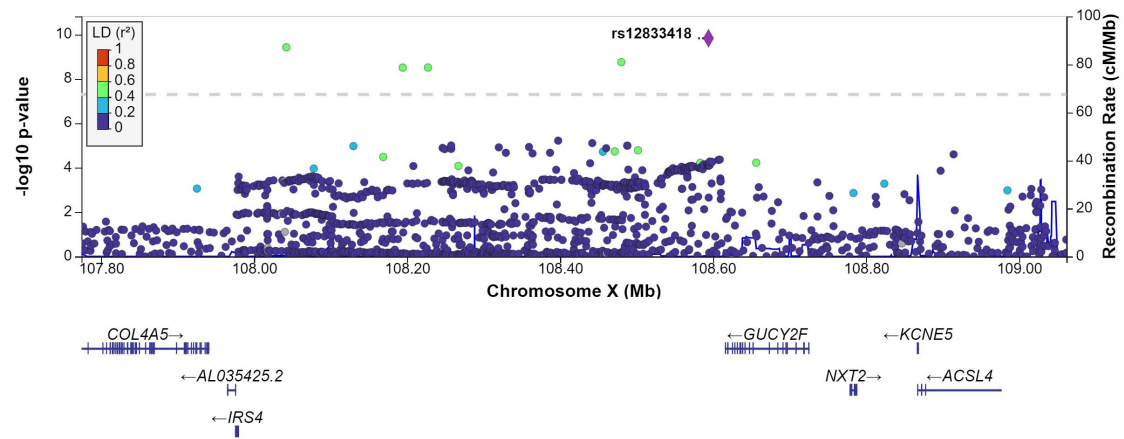
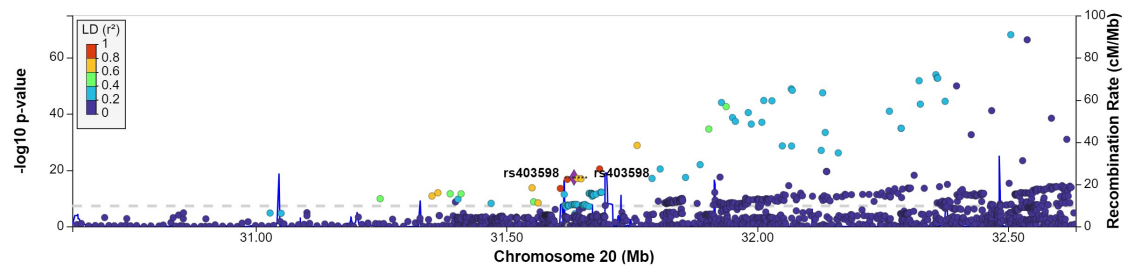
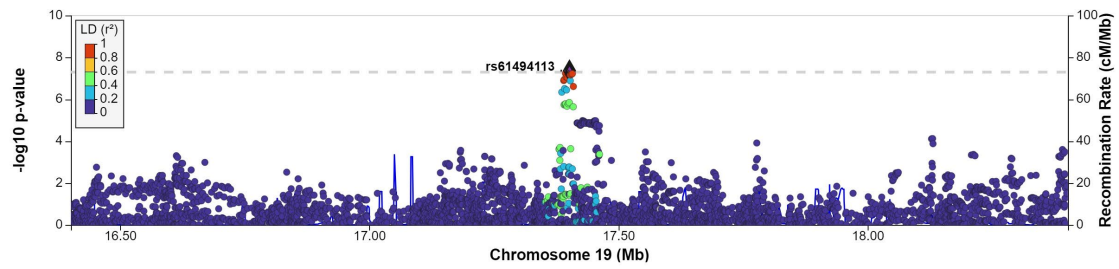




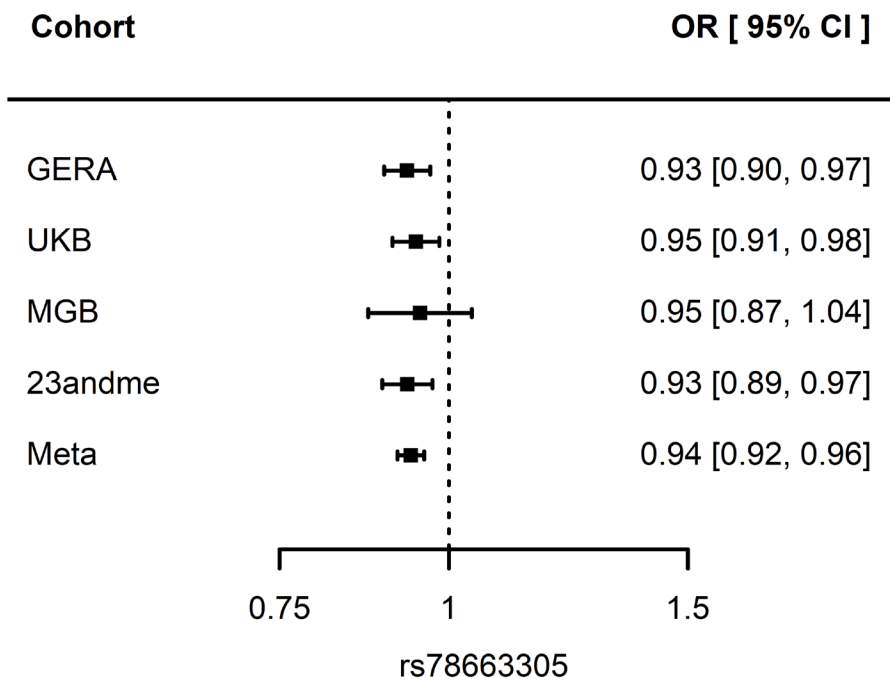
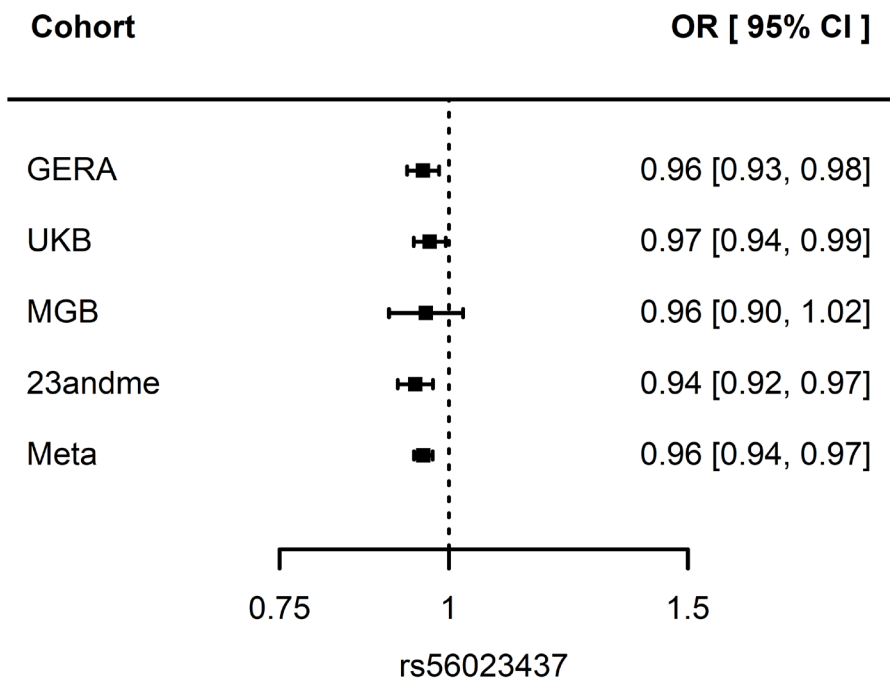


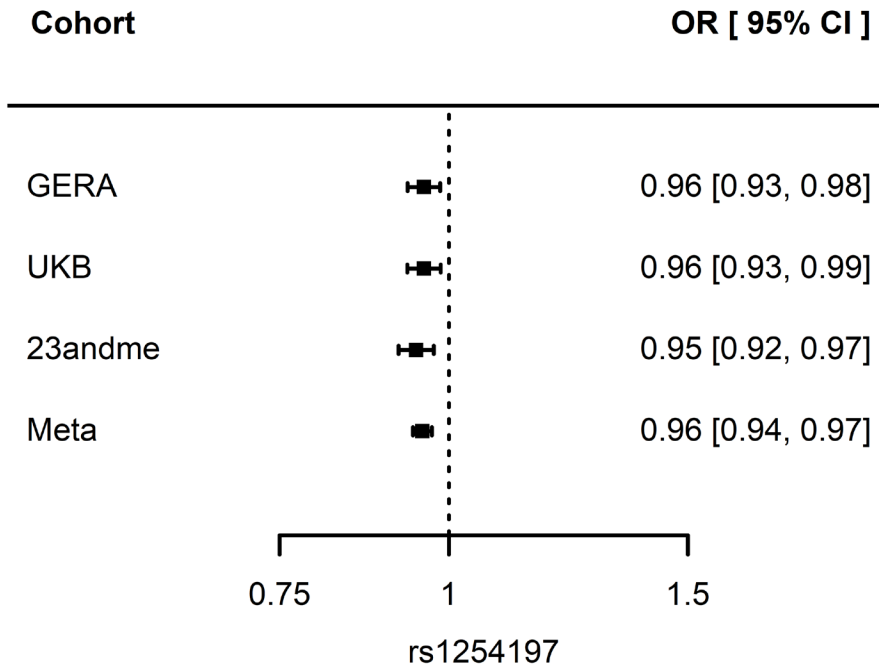
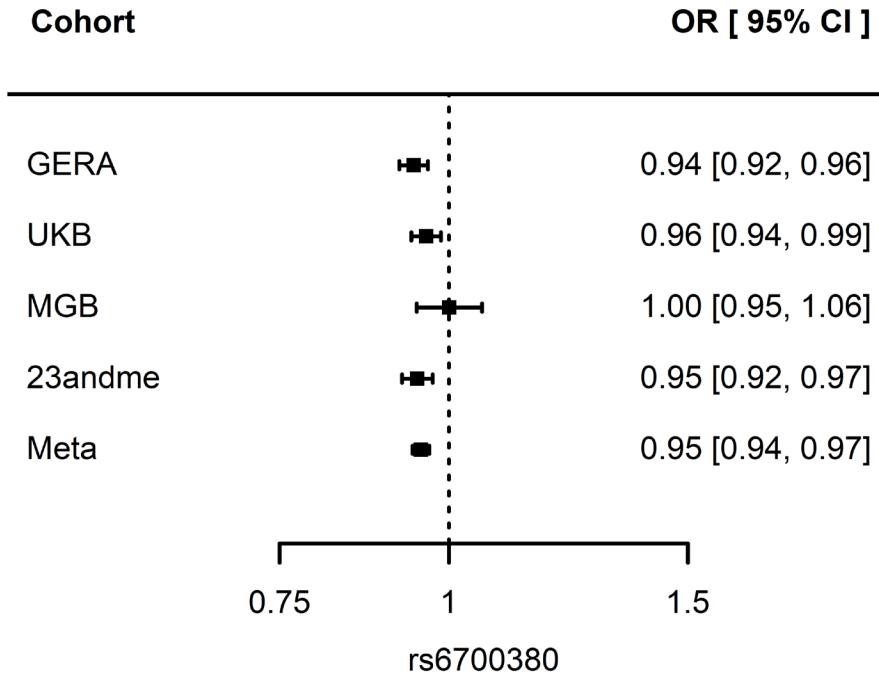


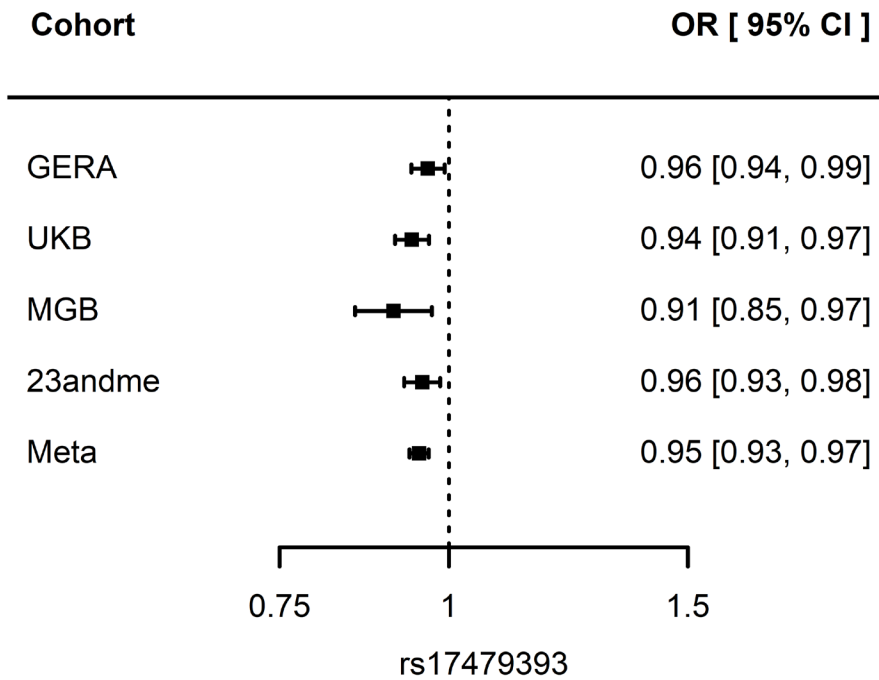
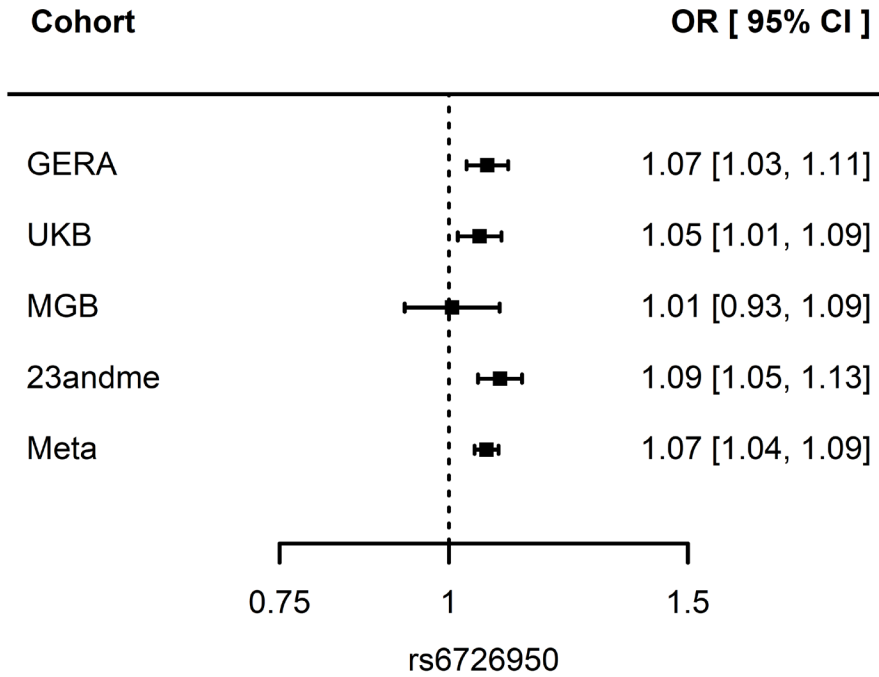


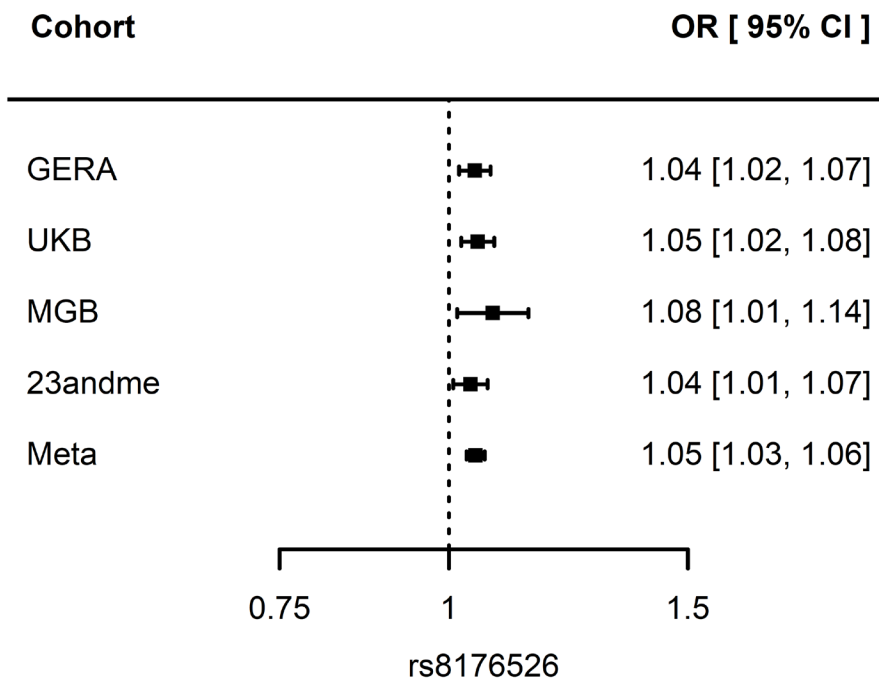
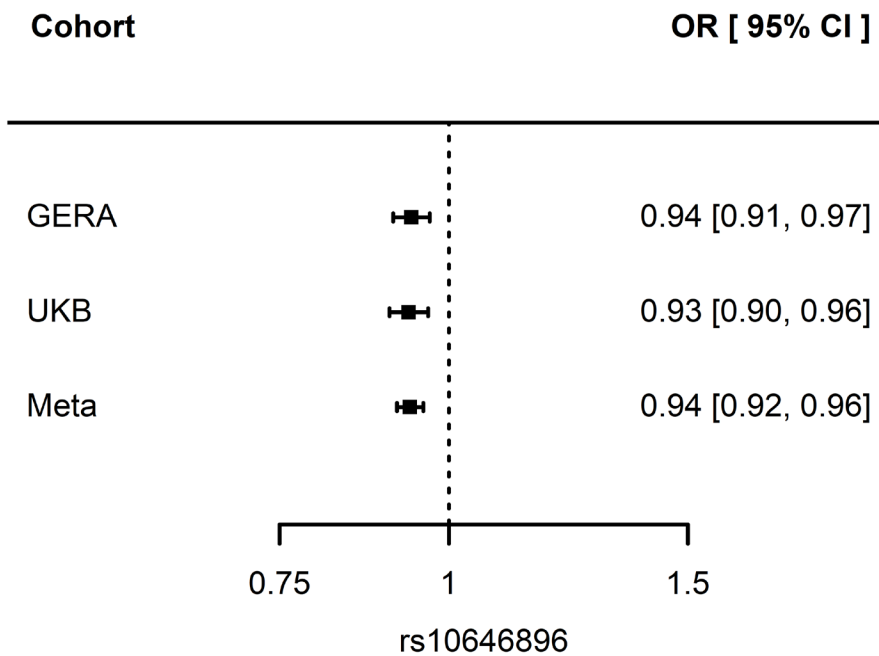


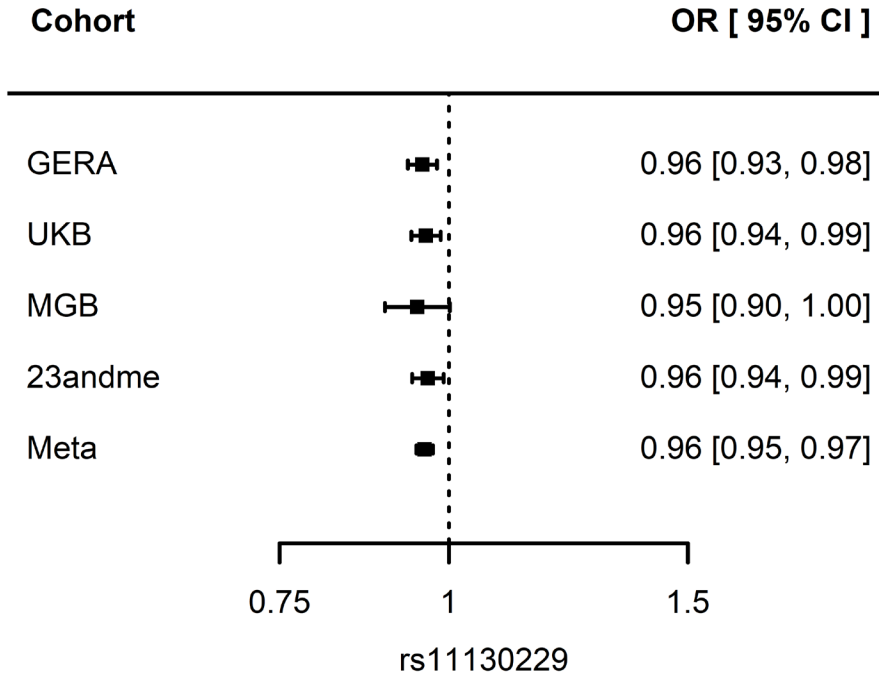
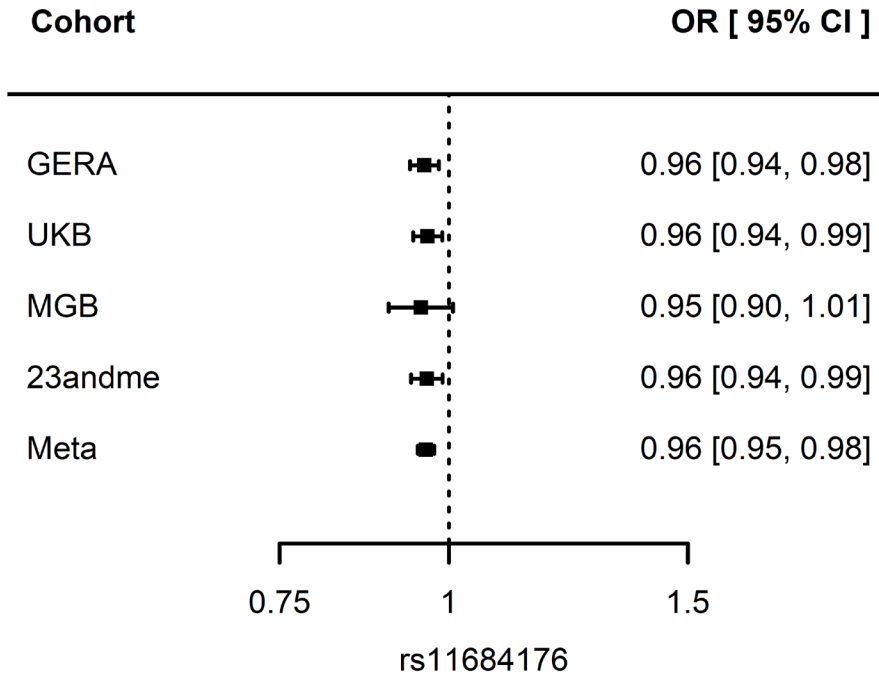
Supplementary Figure 4. Forest plots of new BCC loci identified in the European ancestry GWA meta-analysis. The error bars correspond to the 95% CIs.

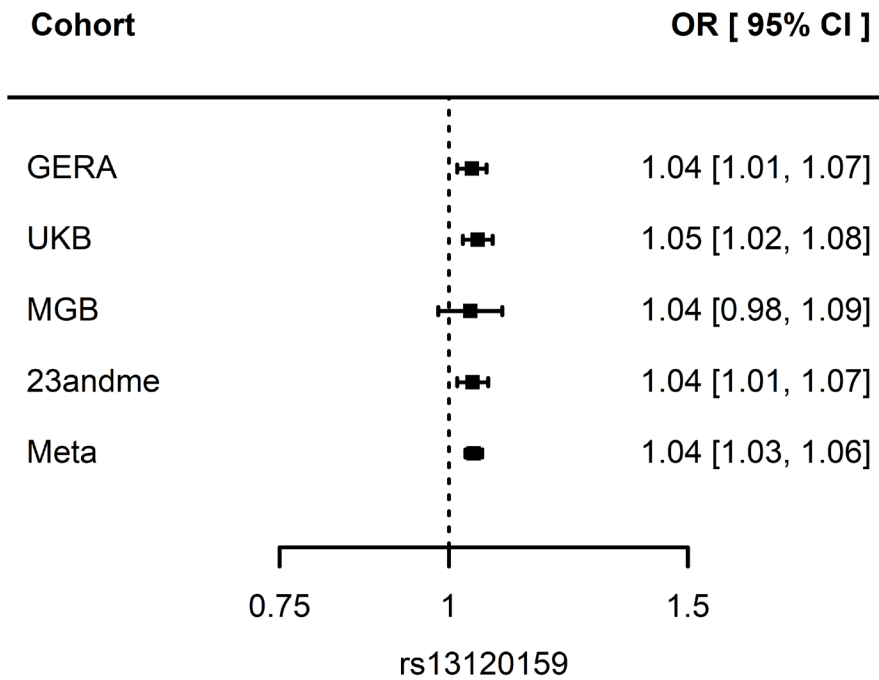
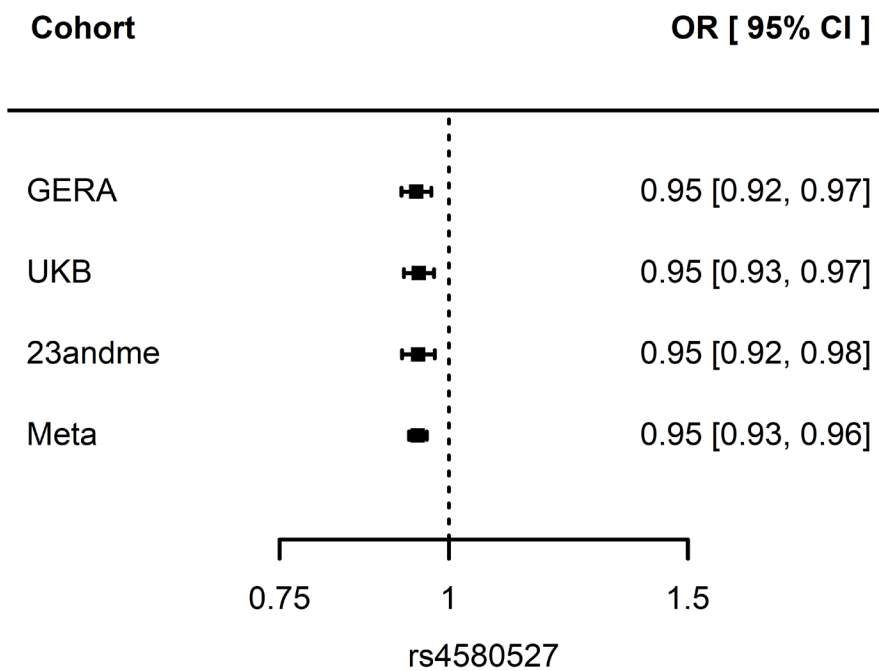


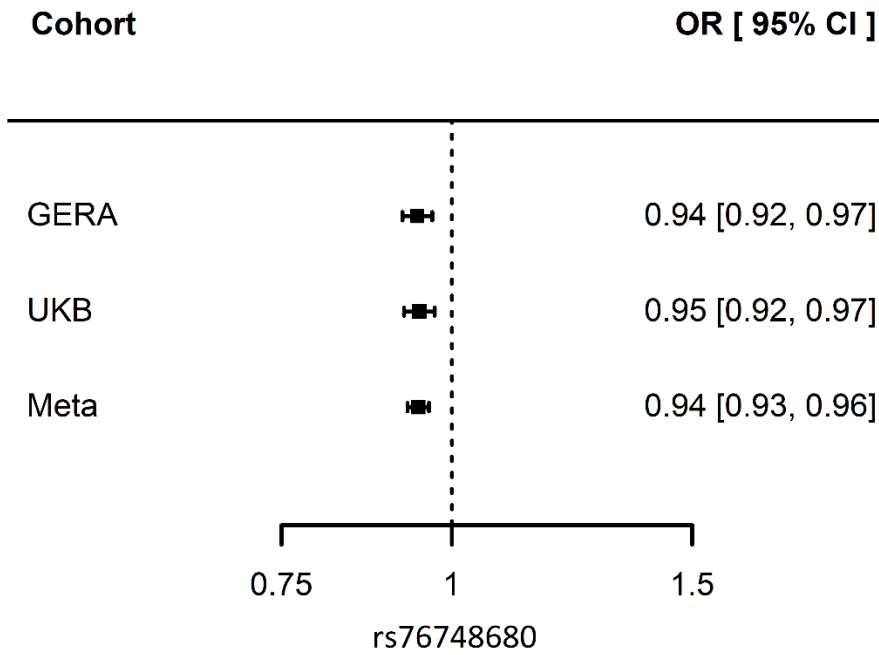
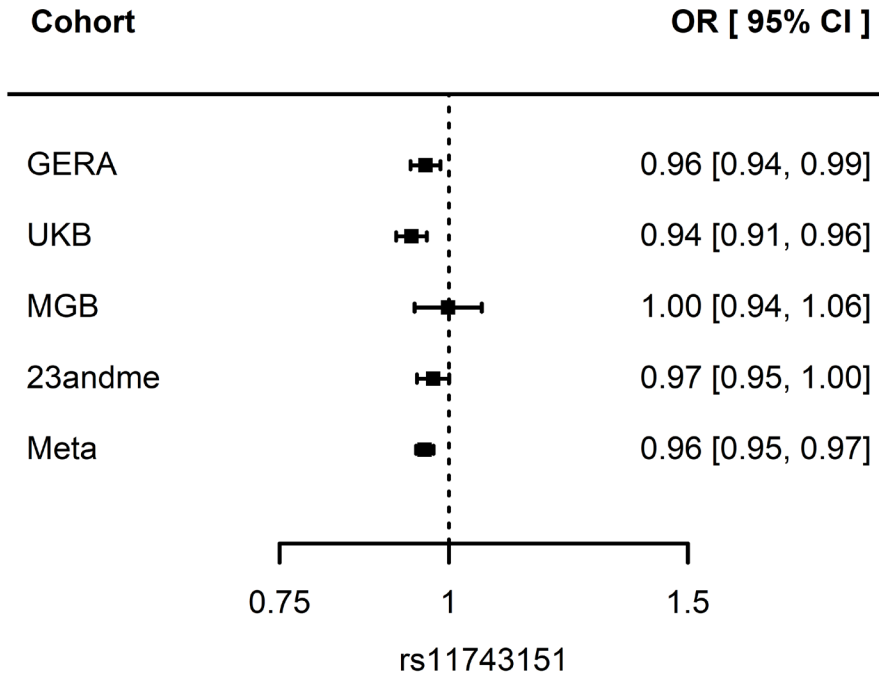


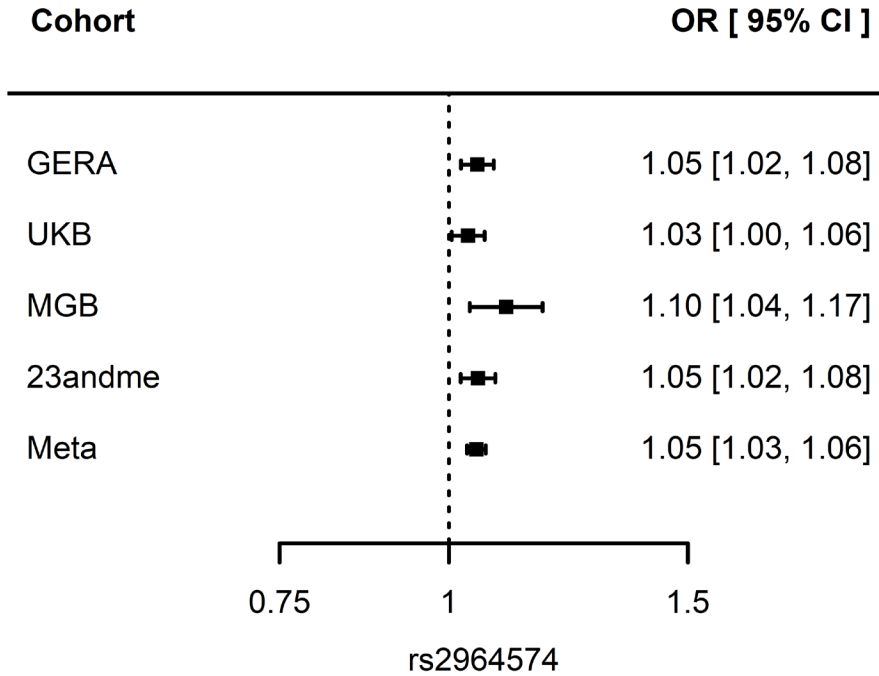
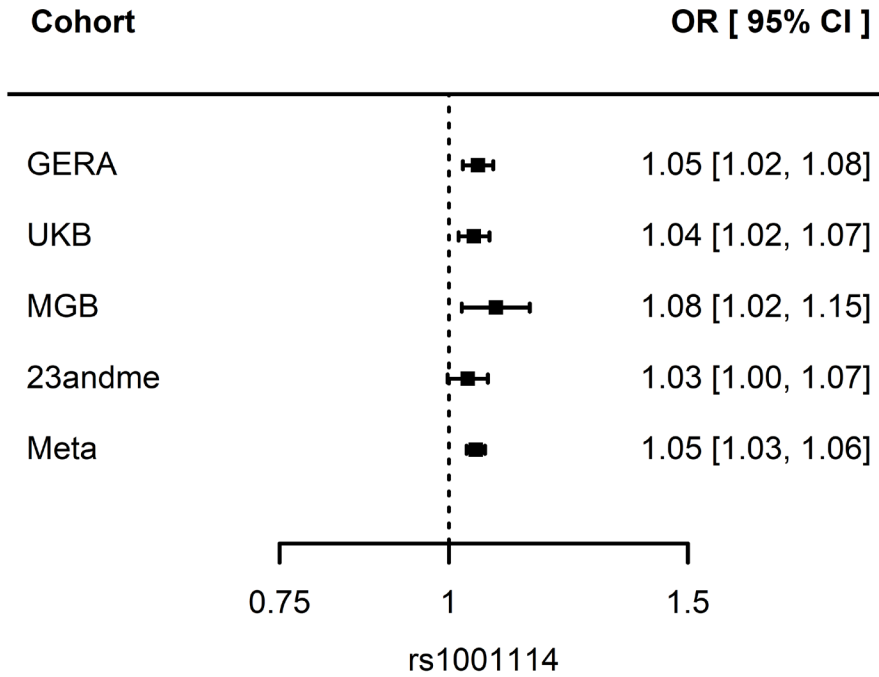


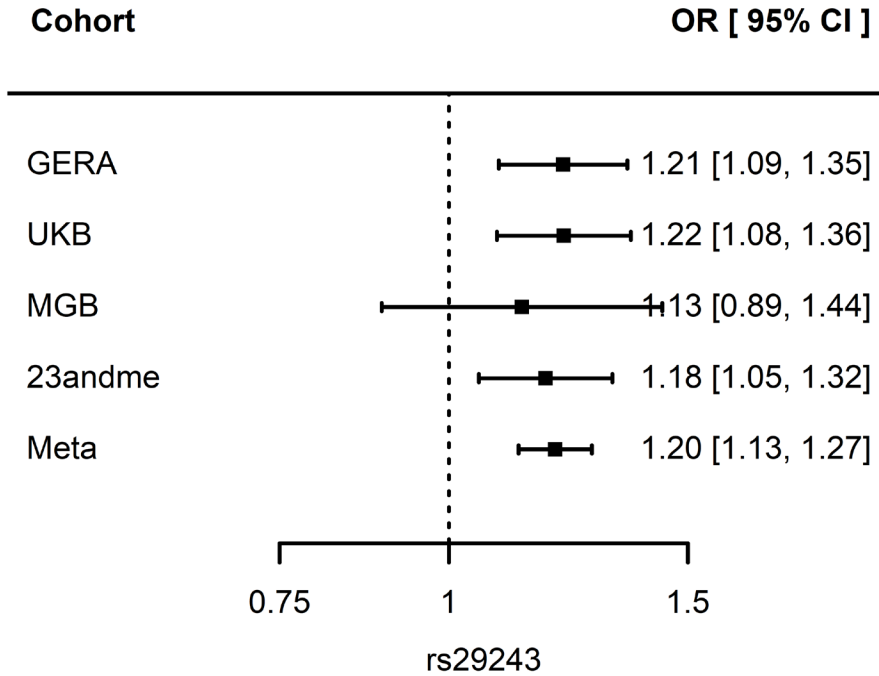
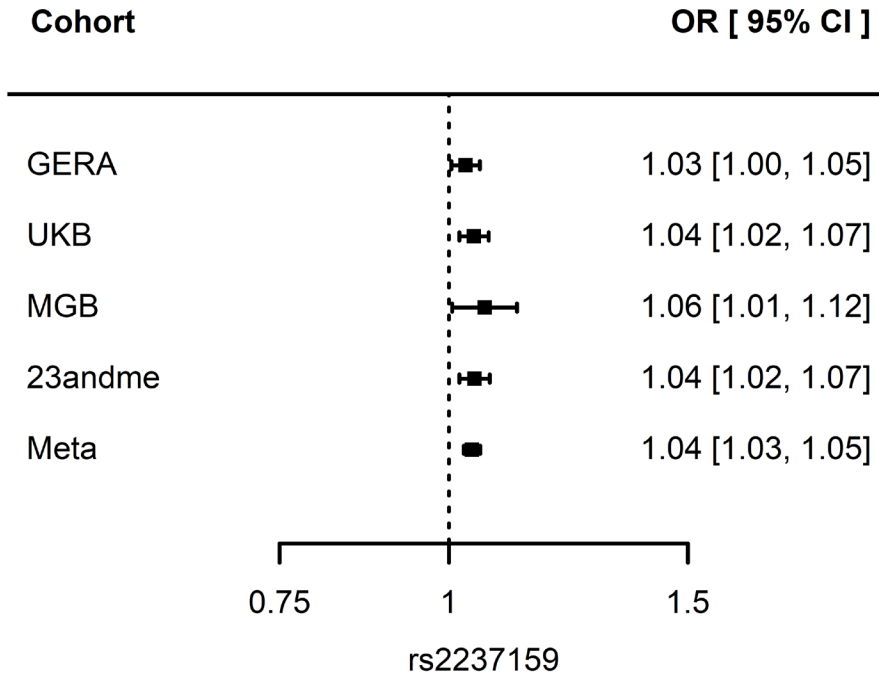


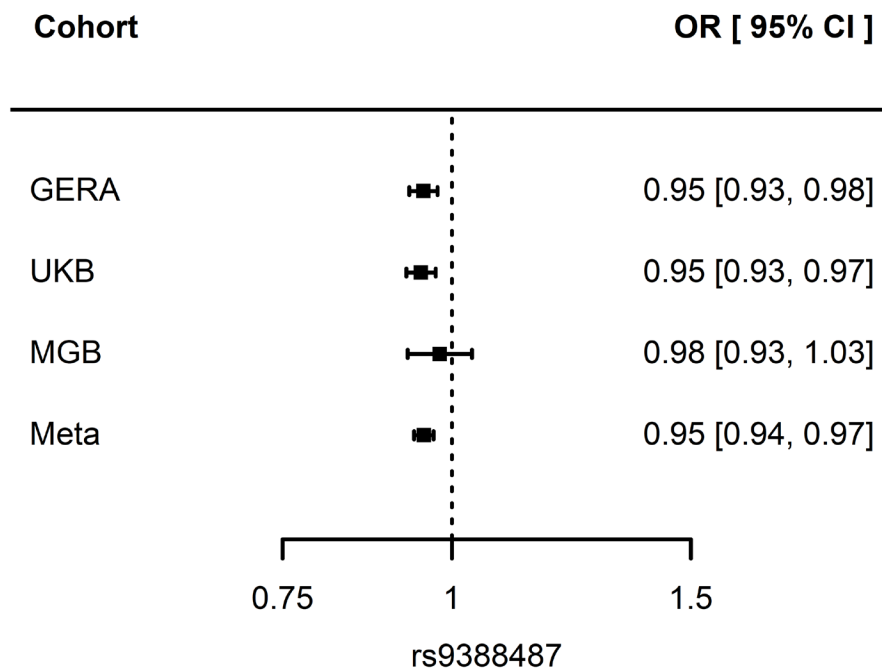
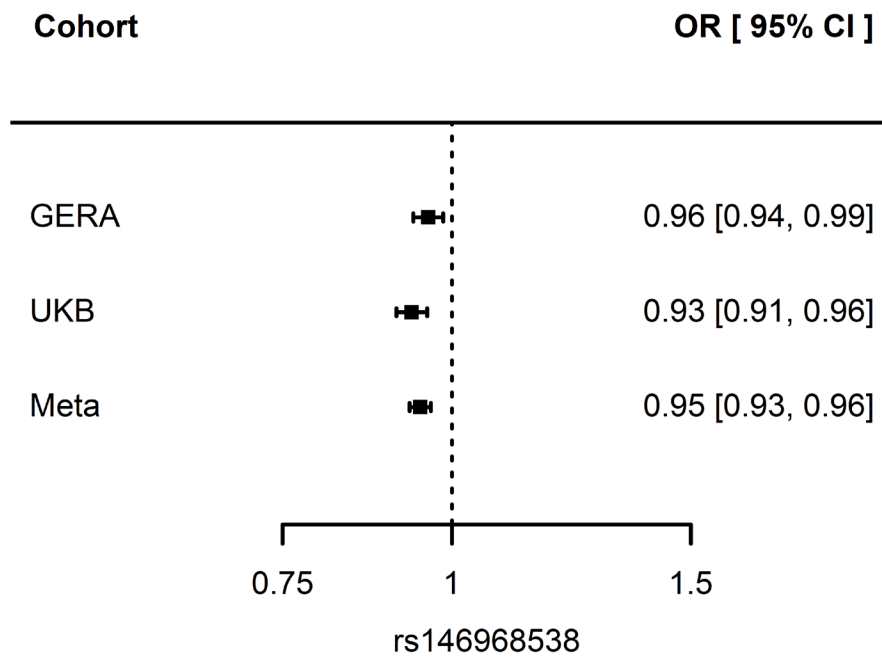


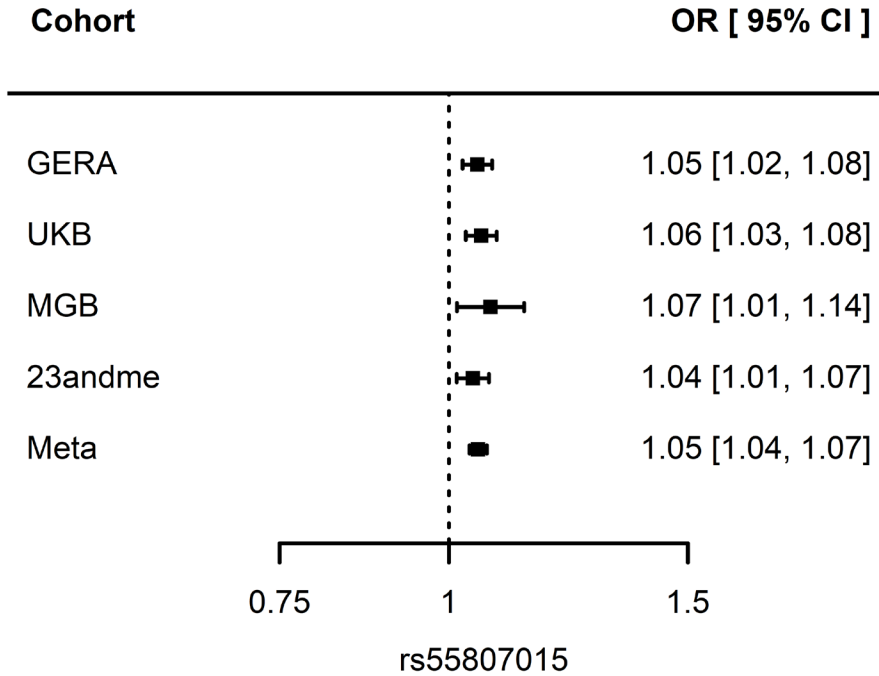
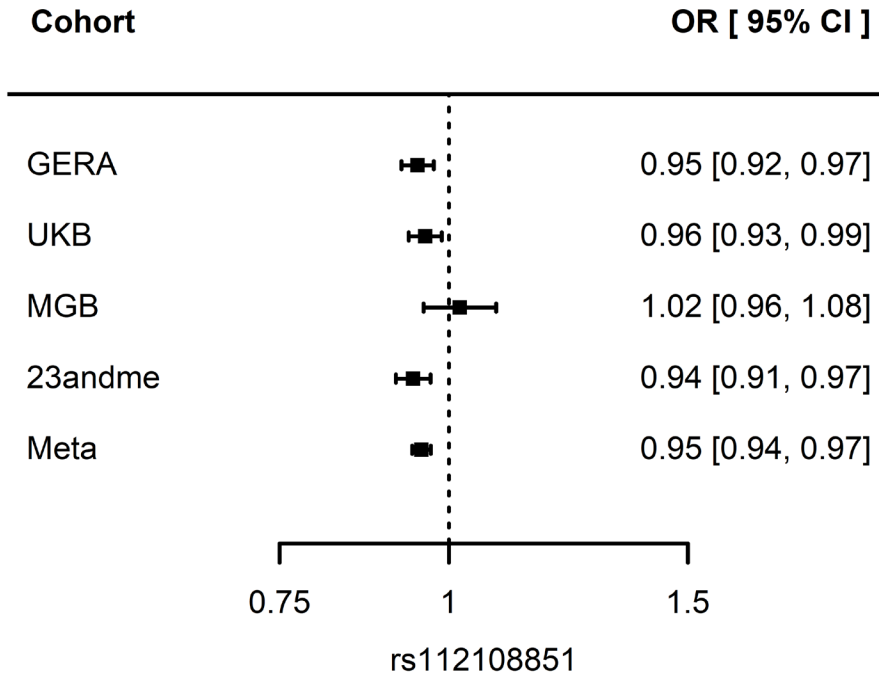


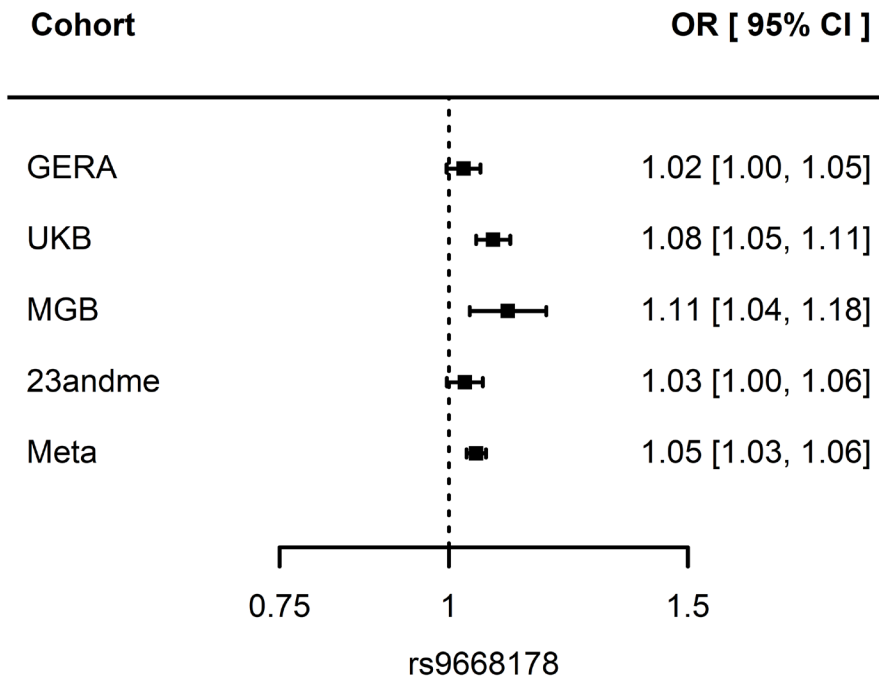
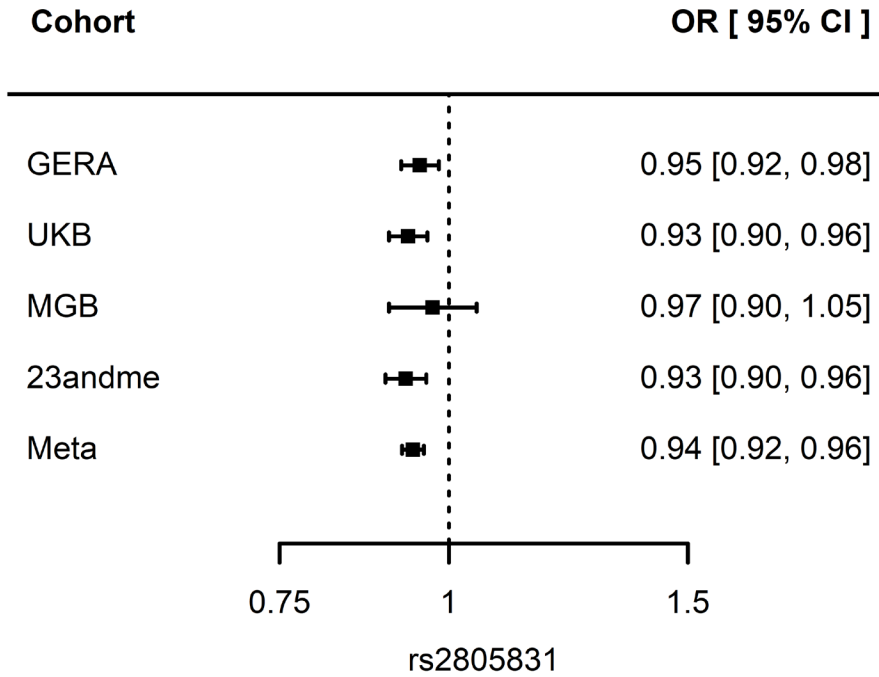


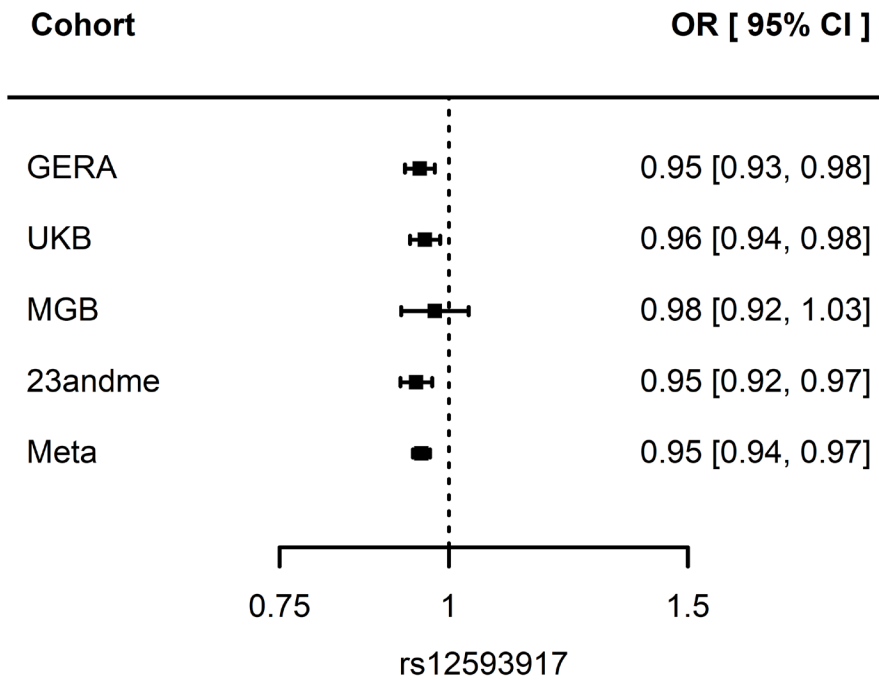
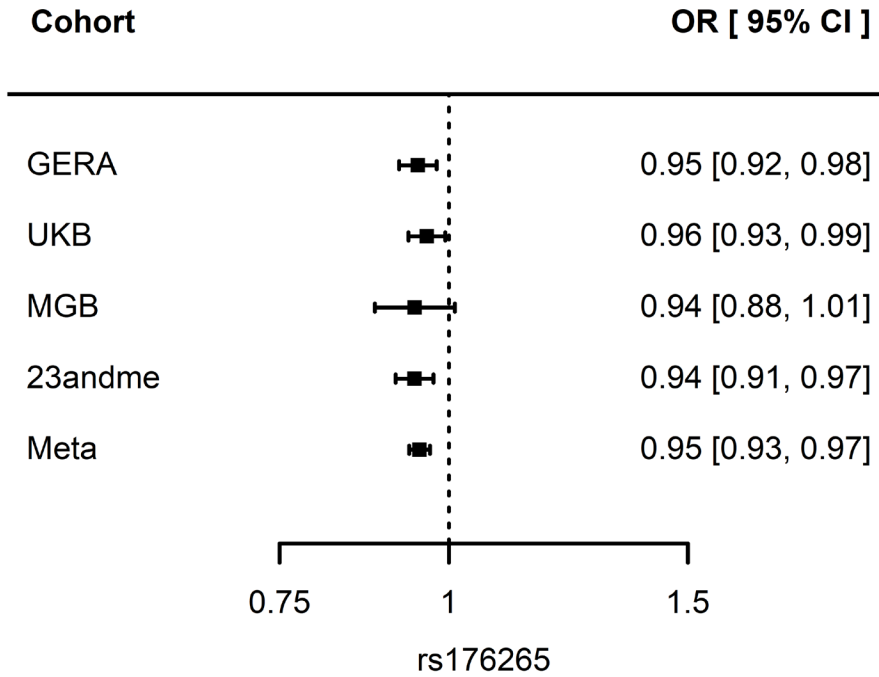


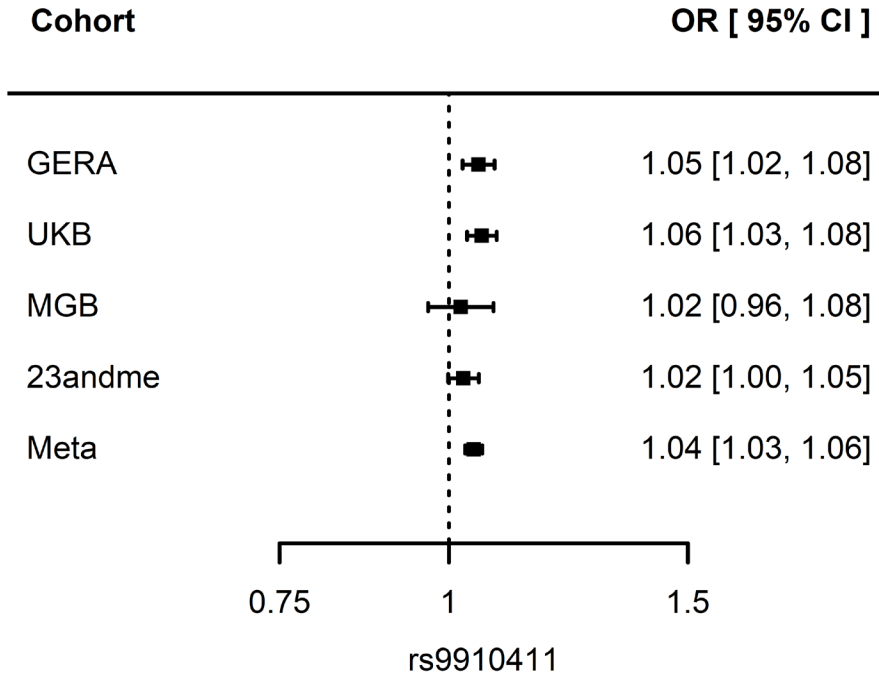
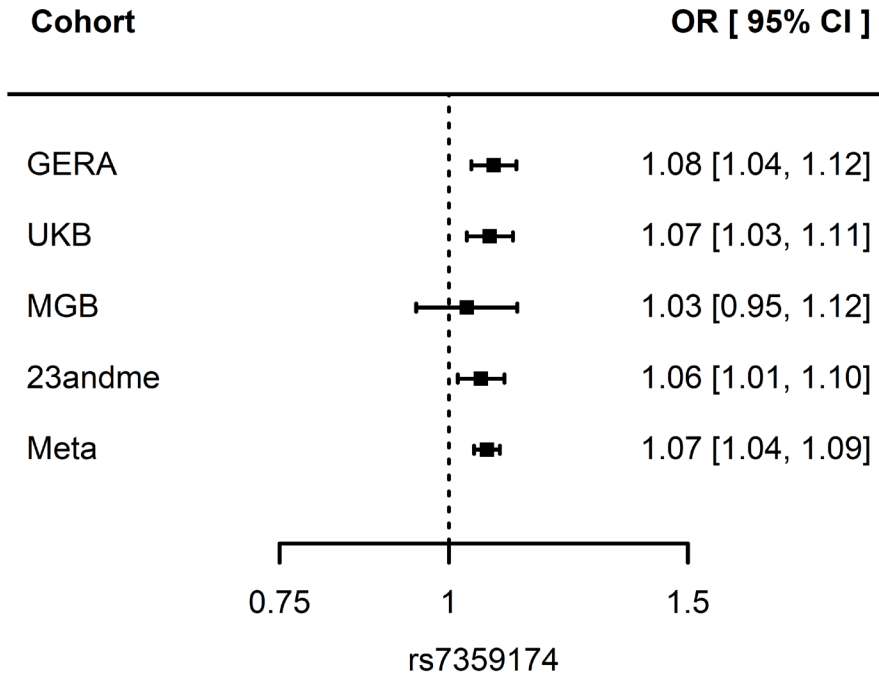


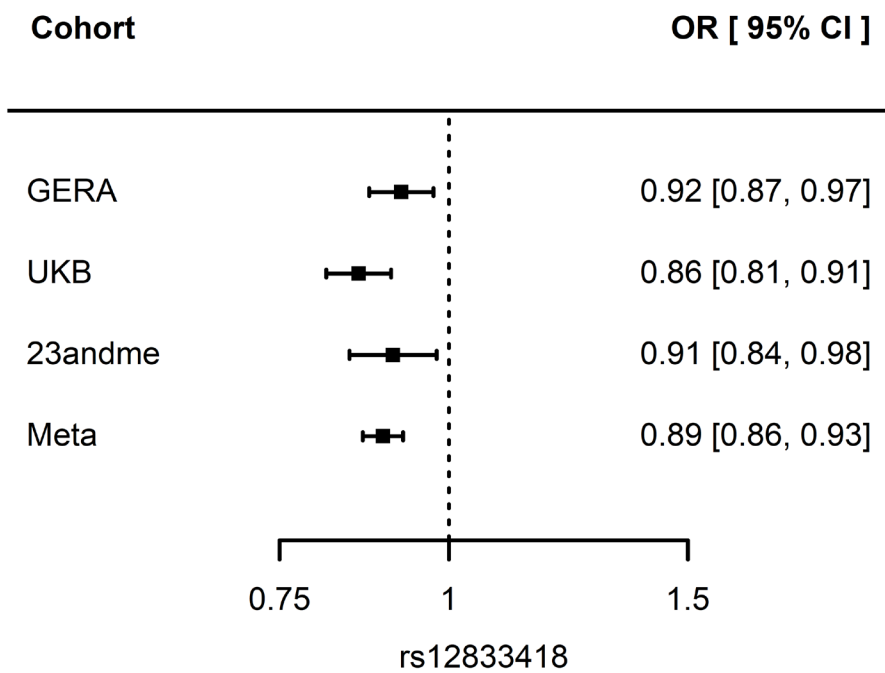
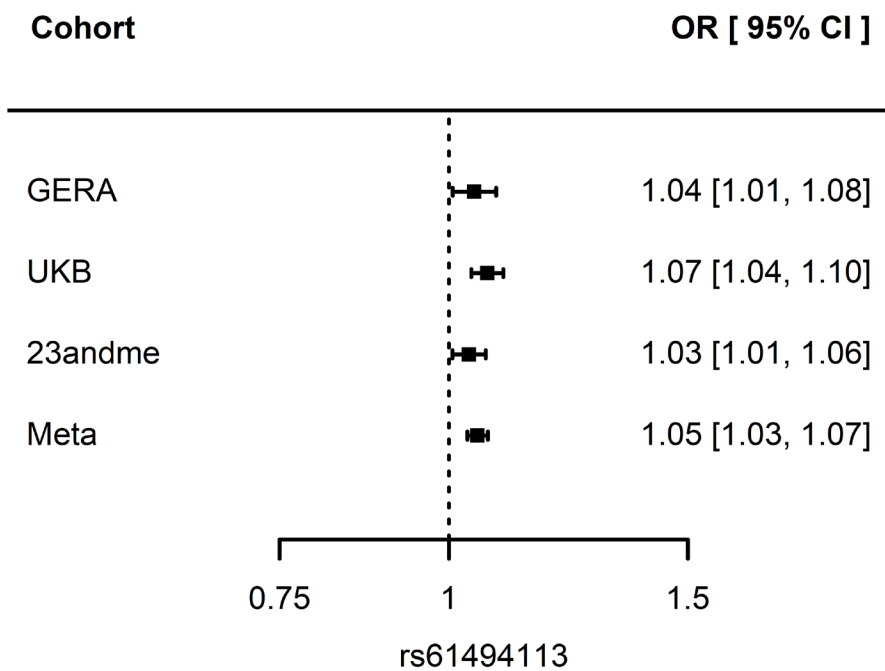




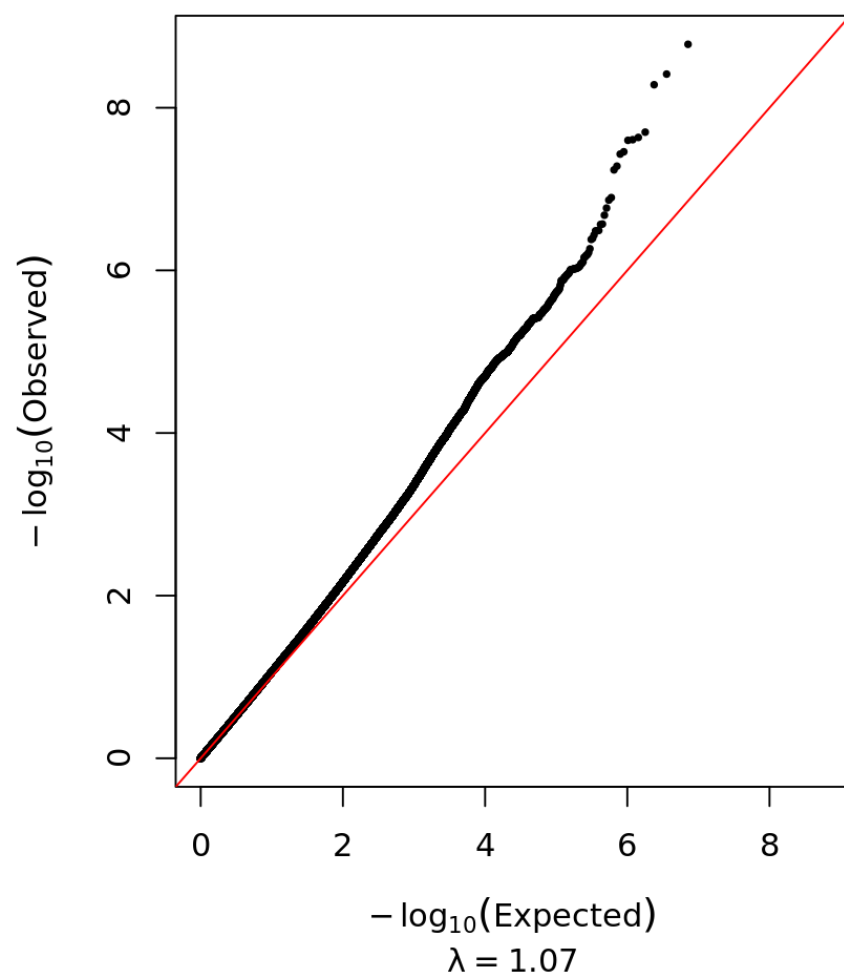




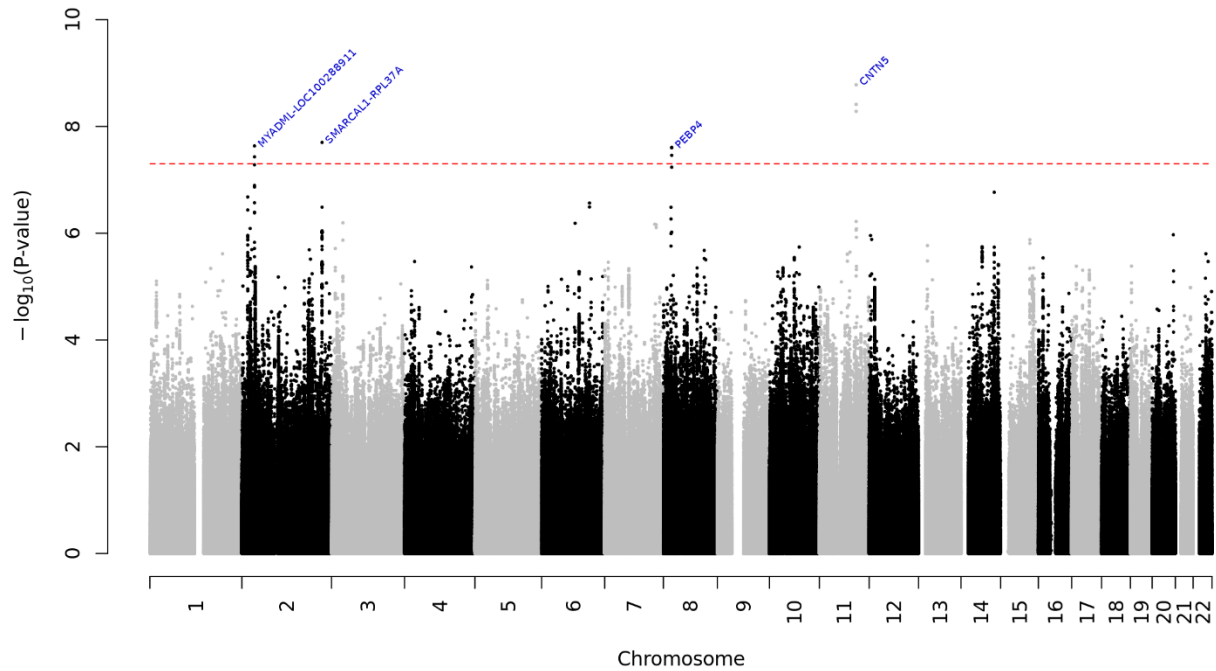




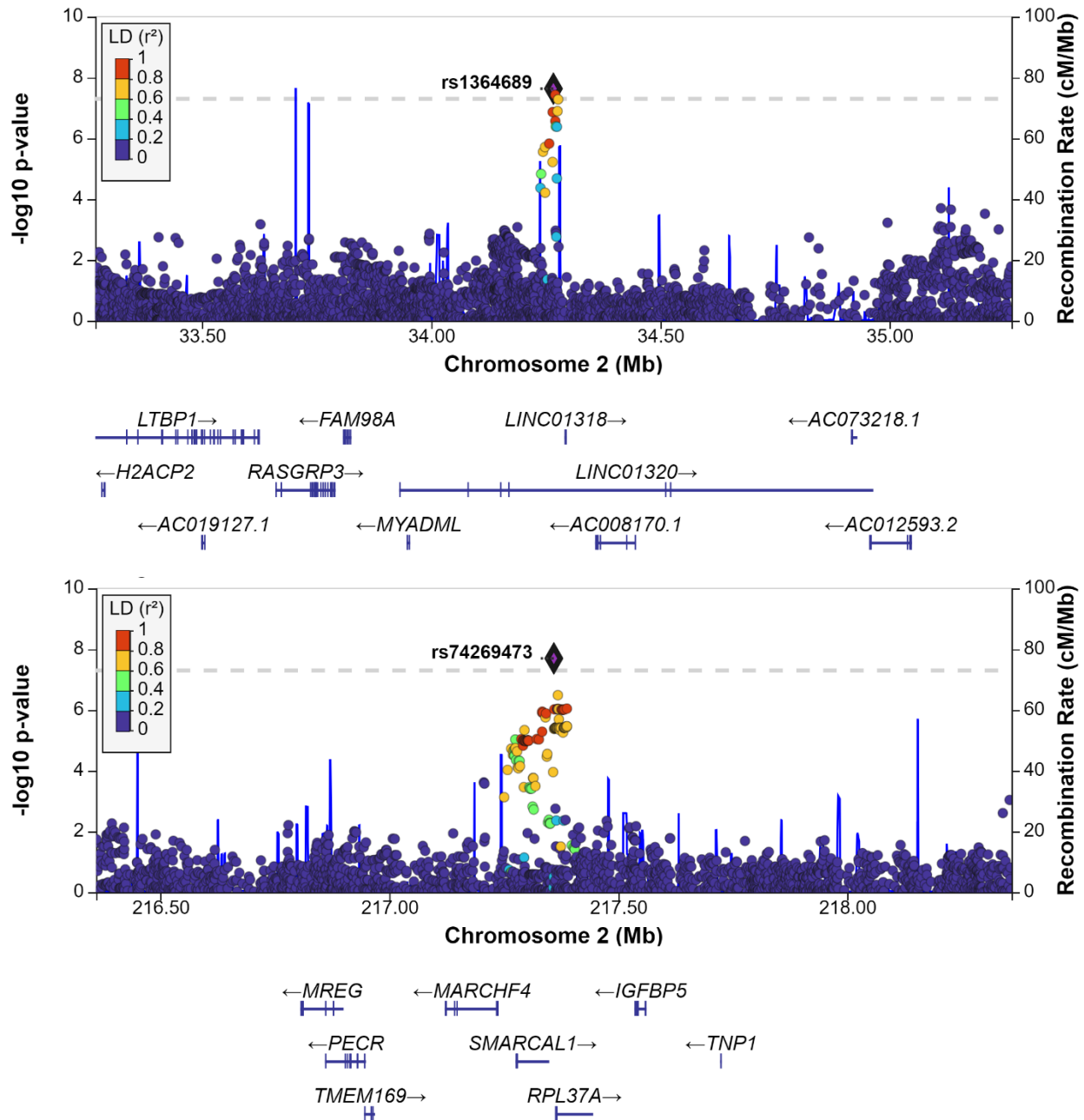
Supplementary Figure 5. QQ plot and genomic inflation factor (λ) observed for the Hispanic/Latino ancestry GWA meta-analysis of BCC

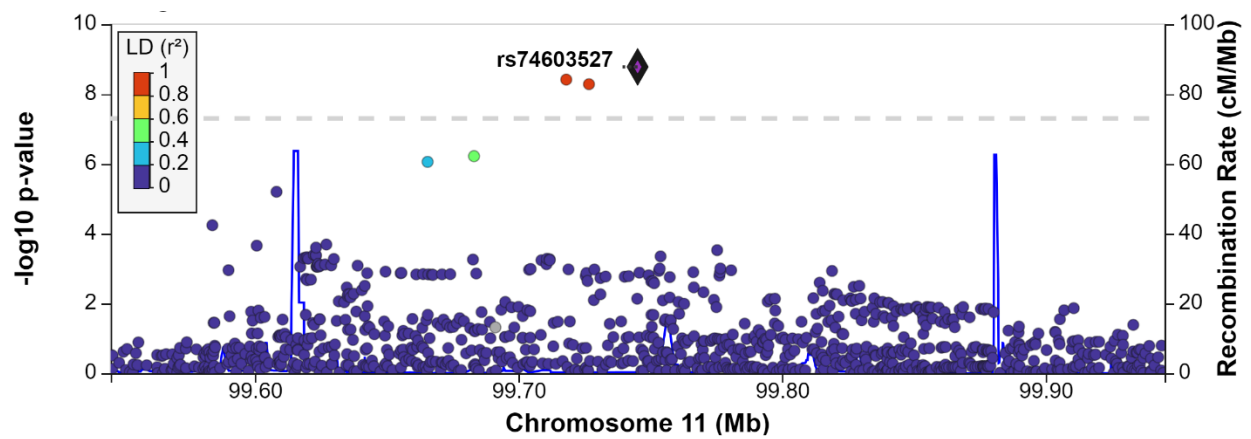
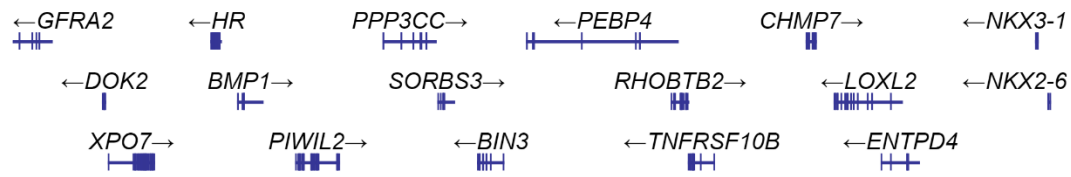
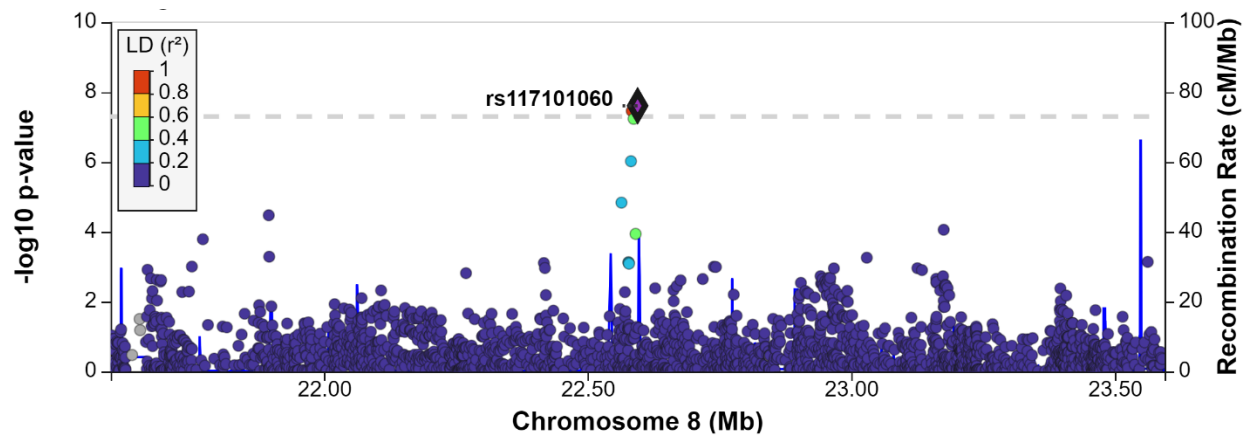


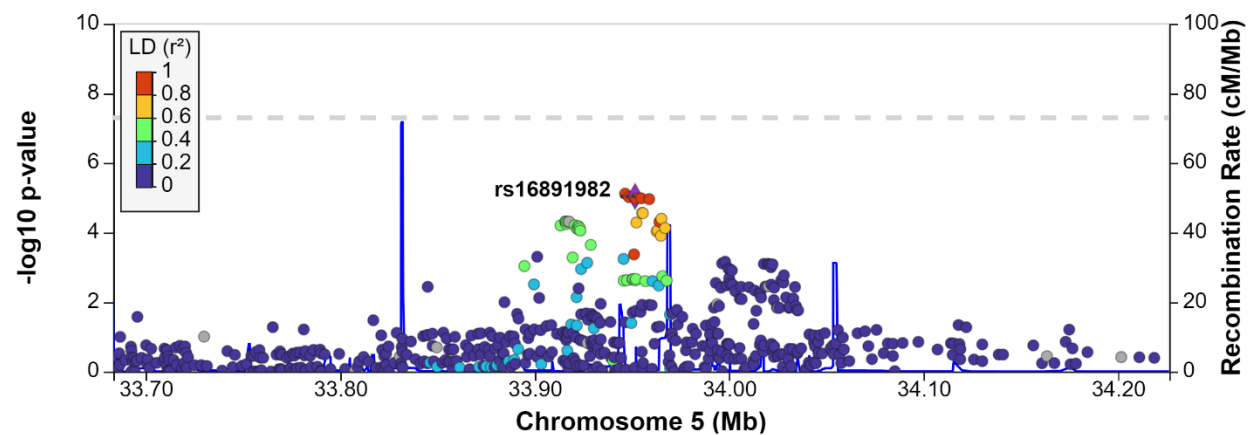
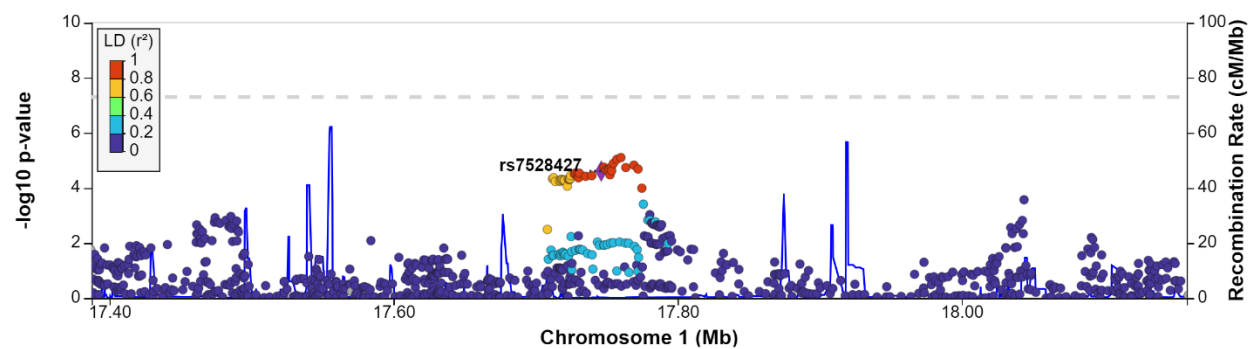
Supplementary Figure 6. Manhattan plot of the Hispanic/Latino ancestry GWA meta-analysis of BCC. The y-axis represents the $-\log_{10}(P\text{-value})$; all P -values derived from logistic regression model are two-sided. The red dotted line represents the threshold of $P=5 \times 10^{-8}$ which is the commonly accepted threshold of adjustments for multiple comparisons in GWAS.

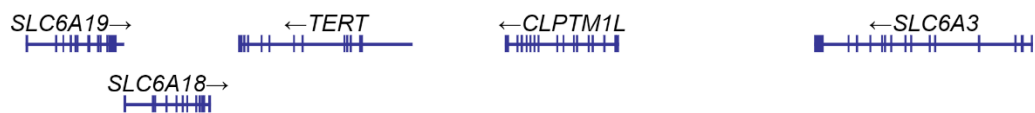
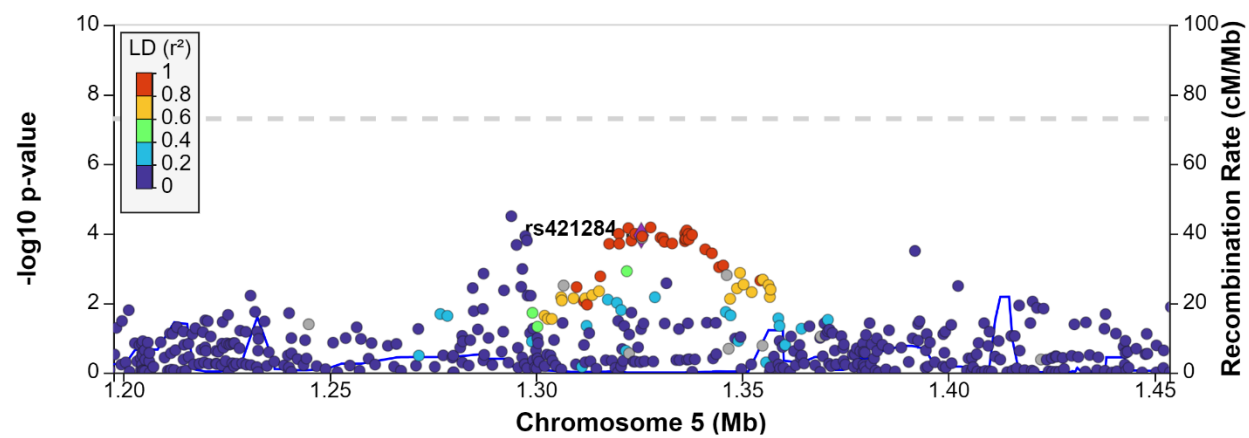


Supplementary Figure 7. Locus Zoom plots of regions identified in the Hispanic/Latino GWA meta-analysis of BCC or showing association with BCC at Bonferroni level of significance. Results are from the Hispanic/Latino ancestry GWA meta-analysis combining data from the GERA and MGB cohorts.

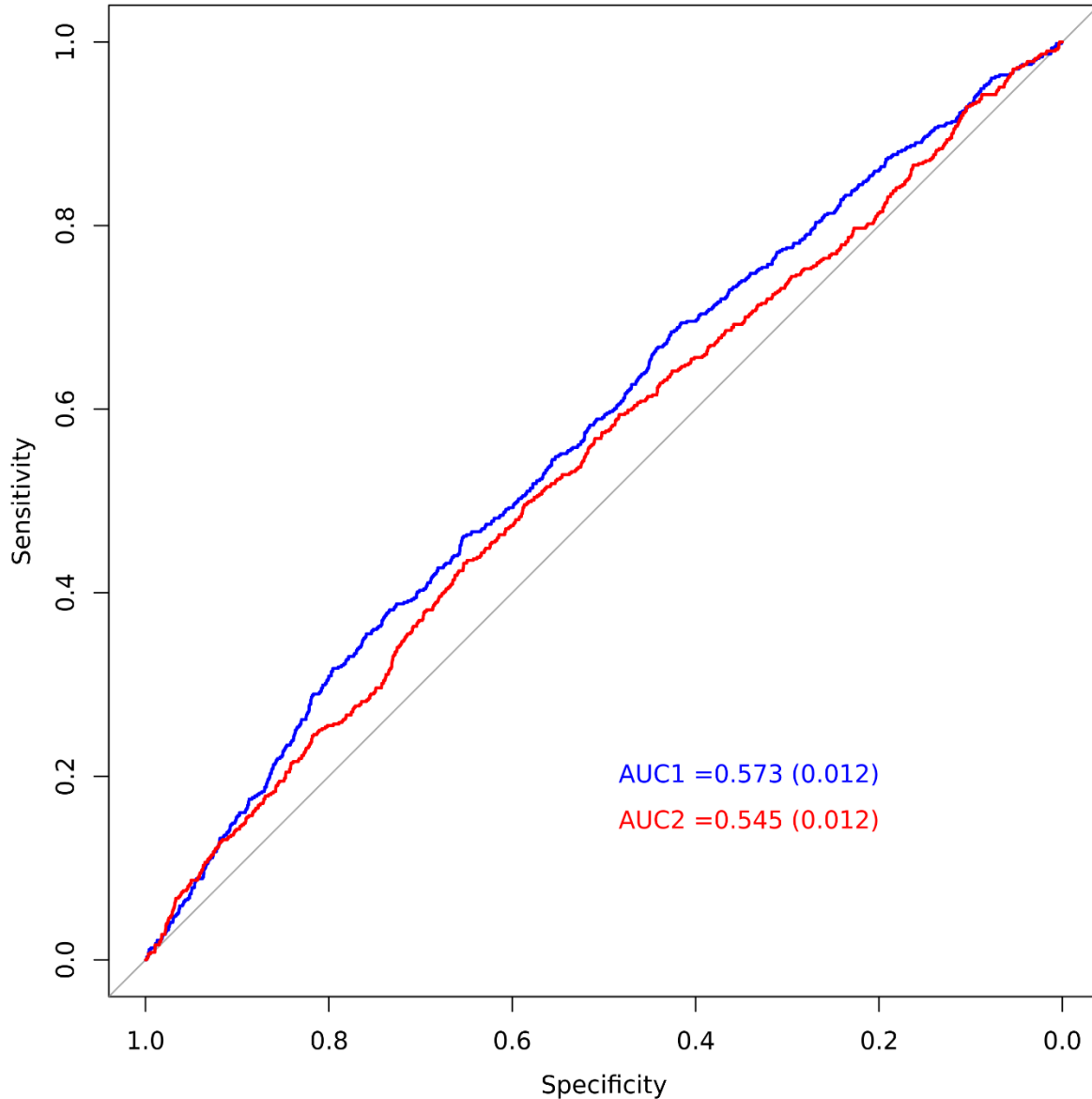




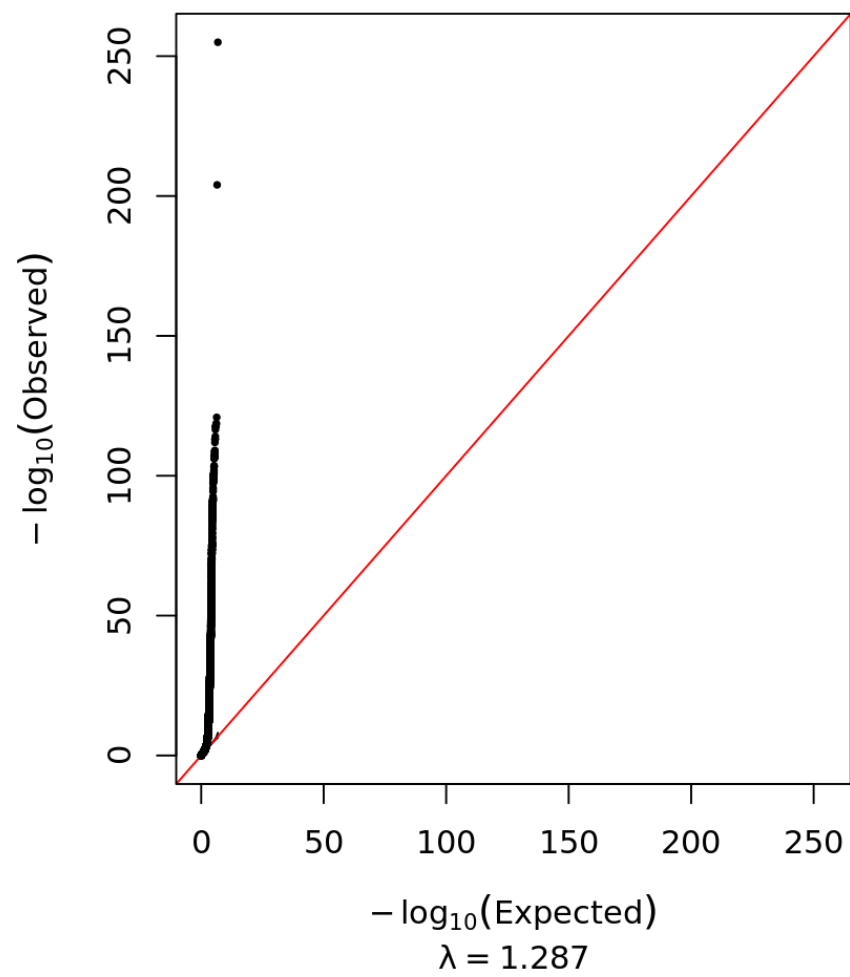




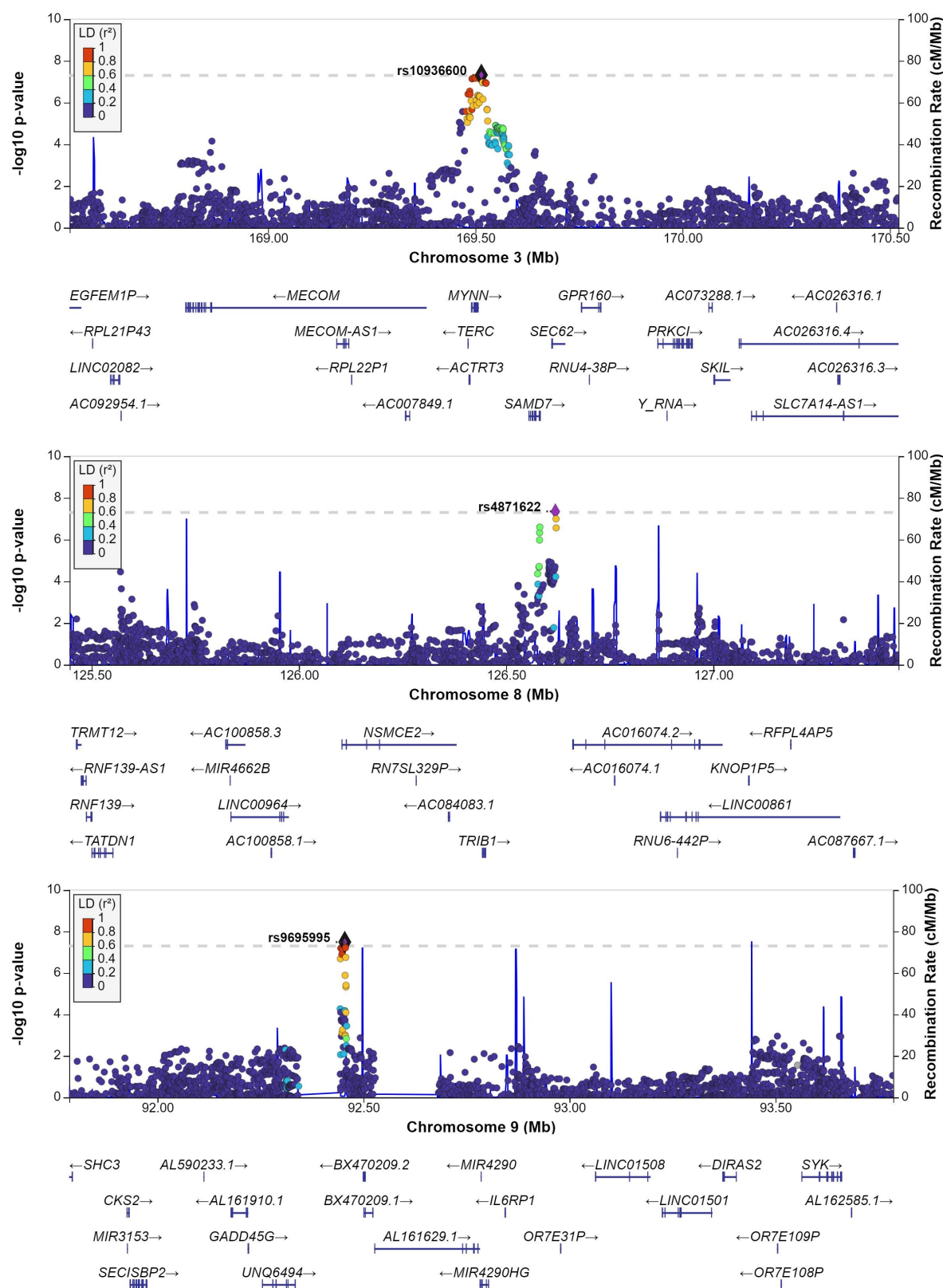
Supplementary Figure 8. Receiver operating characteristic curves for BCC predictions in the GERA Hispanic/Latino sample, using polygenic risk scores derived from the European ancestry GWA meta-analysis of BCC. The two different colors represent two different curves for each of the different EUR-derived PRSs of BCC: blue, PRS included clumped independent SNPs at P -value threshold $< 5.0 \times 10^{-8}$; and red, PRS included clumped independent SNPs at P -value threshold $< 1.0 \times 10^{-6}$. AUC values and standard deviations are reported to the legend. Abbreviation: AUC, area under the curve.



Supplementary Figure 9. QQ plot and genomic inflation factor (λ) observed for the multi-ancestry GWA meta-analysis of BCC

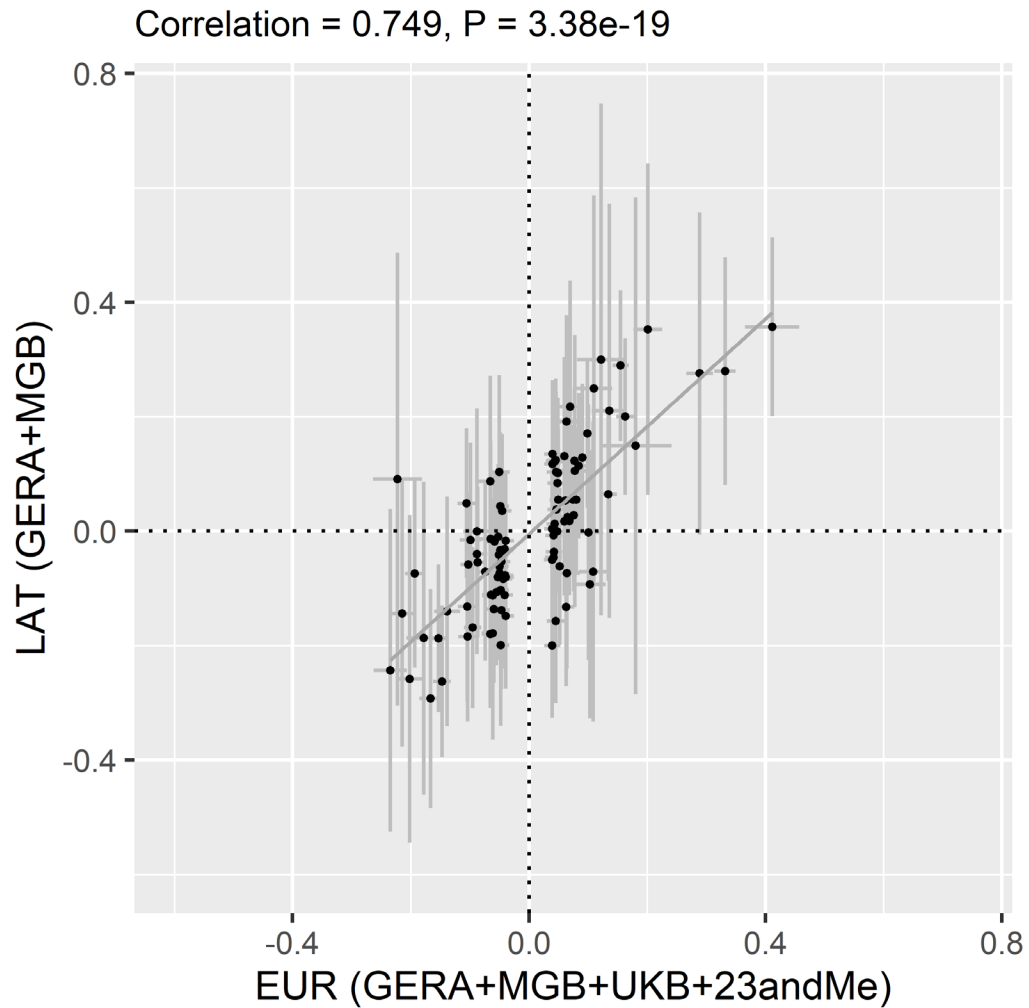


Supplementary Figure 10. Locus Zoom plots of additional novel regions identified in the multi-ancestry GWA meta-analysis of BCC.

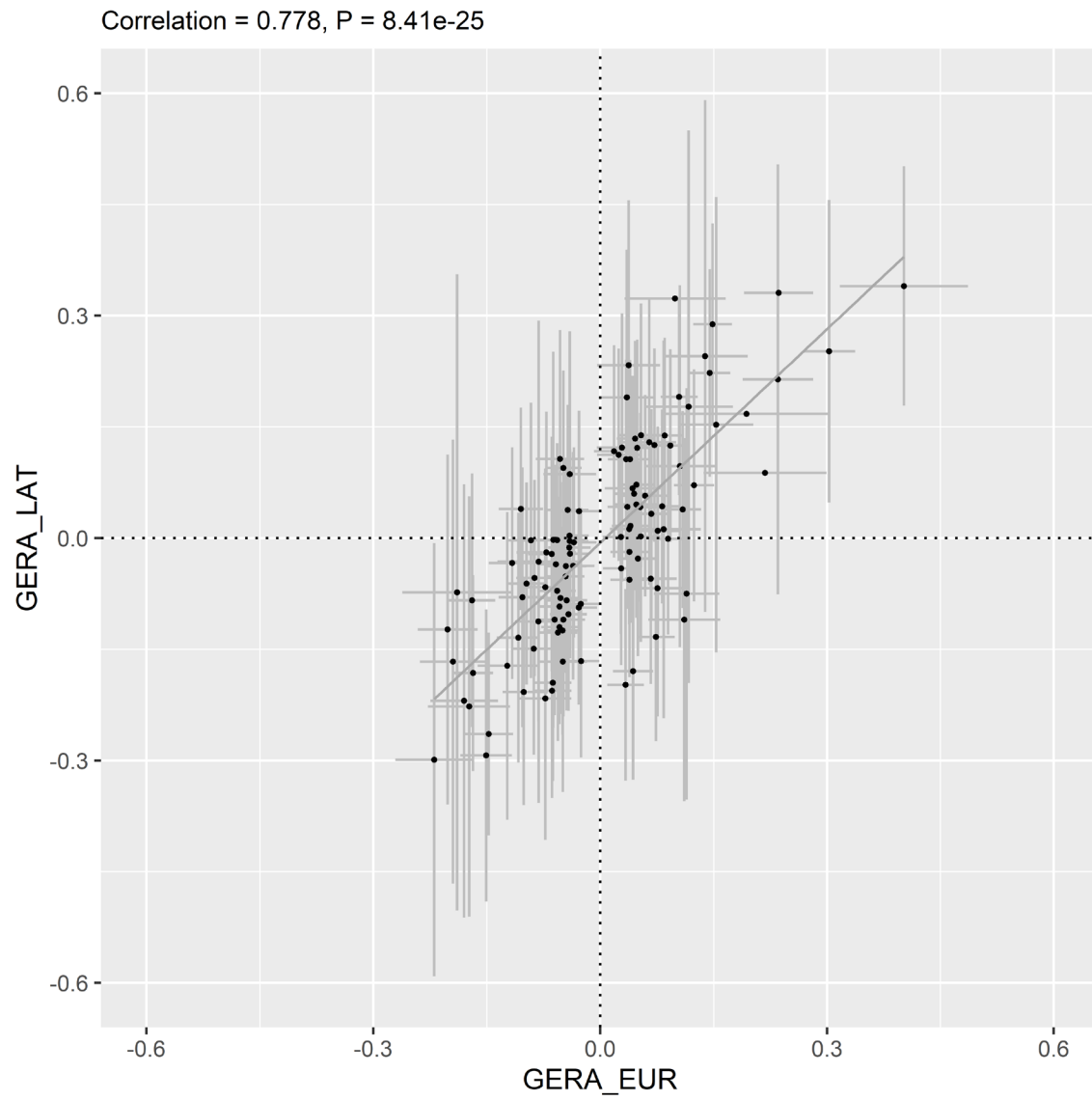


Supplementary Figure 11. Correlation of effect sizes across populations (European ancestry vs. Hispanic/Latino) for the lead 116 BCC-associated lead SNPs identified in the European ancestry GWA meta-analysis of BCC. The error bars correspond to the 95% CIs.

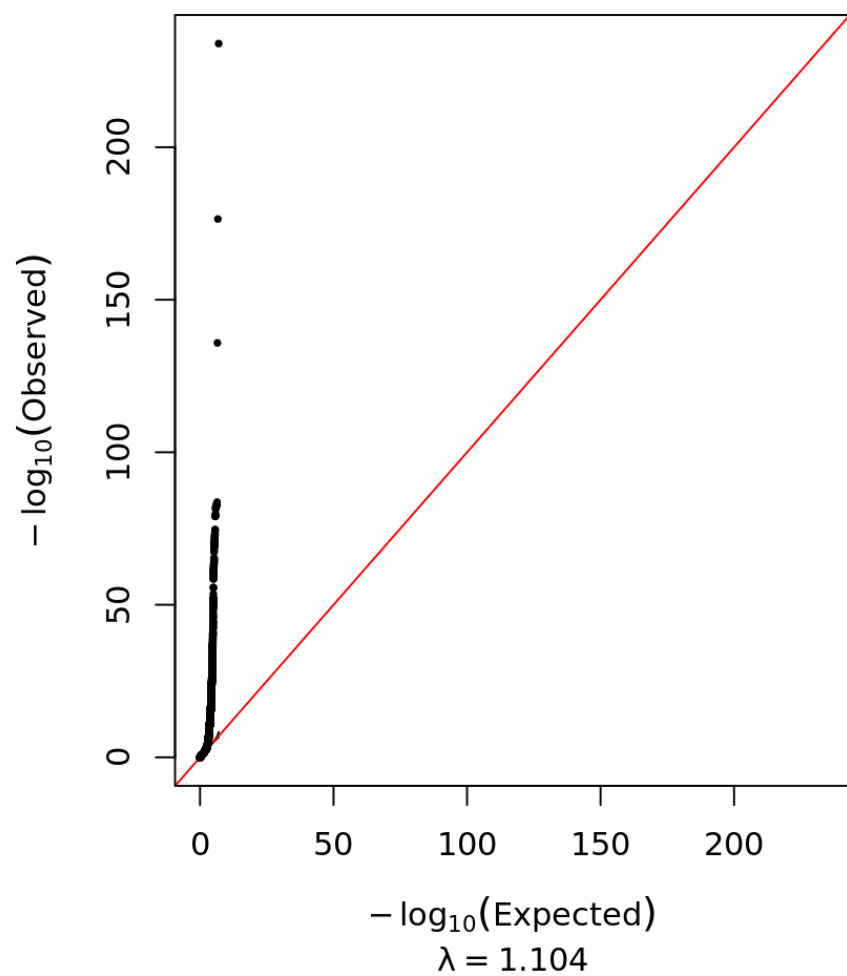
a. EUR (GERA+MGB+UKB+23andMe) vs. LAT (GERA+MGB)



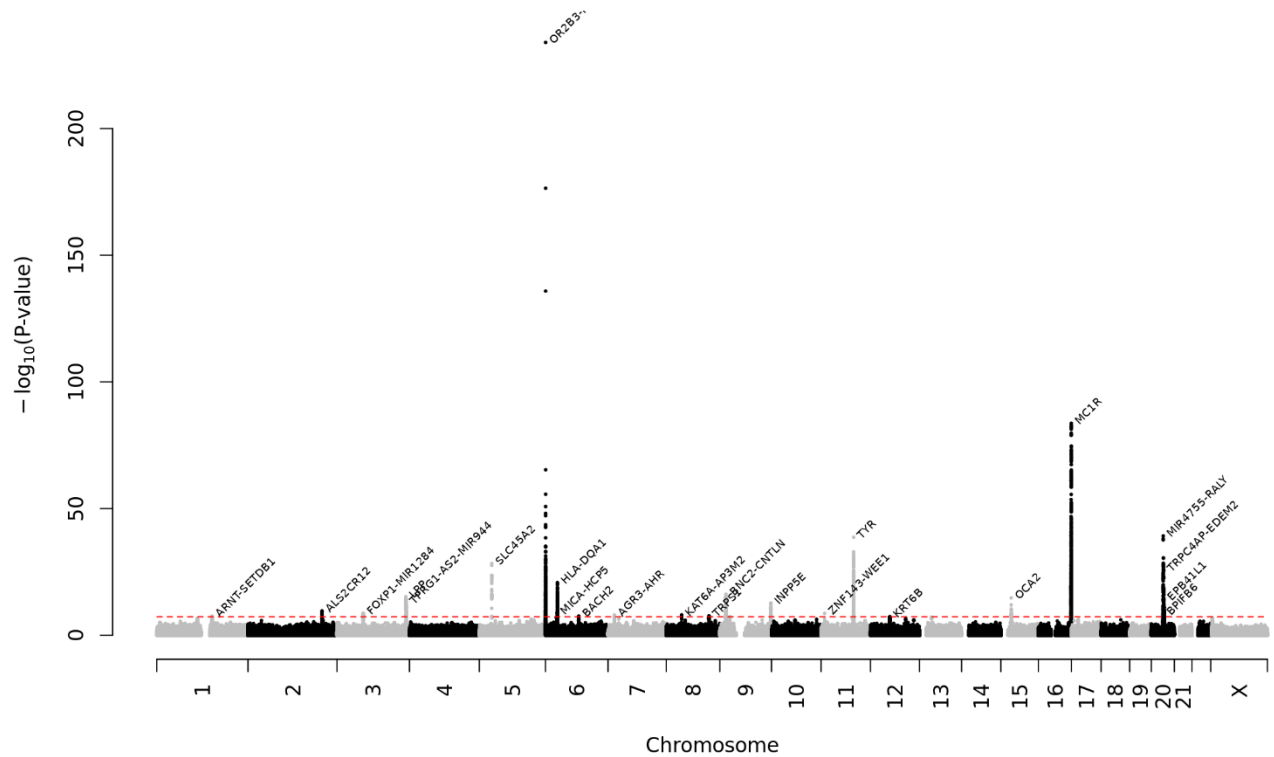
b. EUR (GERA) vs. LAT (GERA)



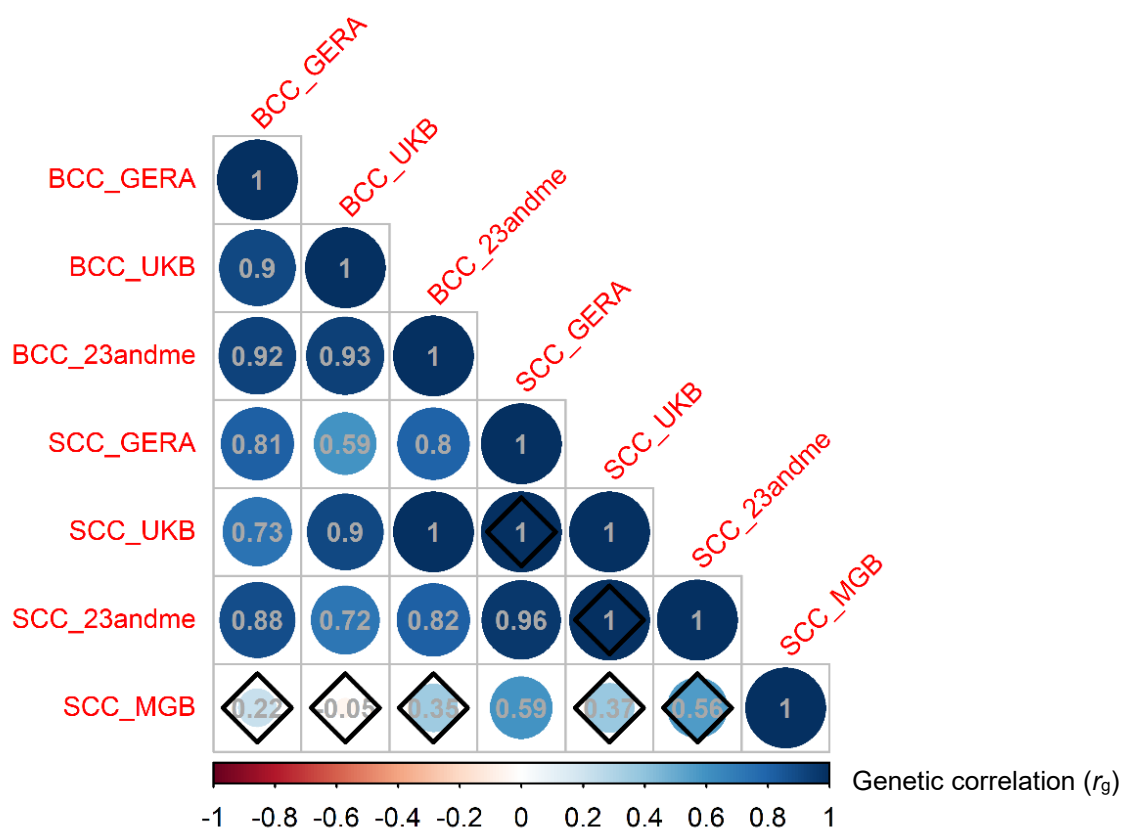
Supplementary Figure 12. QQ plot and genomic inflation factor (λ) observed for the European ancestry GWA meta-analysis of SCC



Supplementary Figure 13. Manhattan plot of the European ancestry GWA meta-analysis of SCC. The y-axis represents the $-\log_{10}(P\text{-value})$; all P -values derived from logistic regression model are two-sided. The red dotted line represents the threshold of $P=5 \times 10^{-8}$ which is the commonly accepted threshold of adjustments for multiple comparisons in GWAS. All loci have been previously reported.



Supplementary Figure 14. Genome-wide genetic correlations between BCC and SCC across cohorts.



Note: summary of genetic correlation results for BCC_MGB are missing, as the input result for the concerned sample has very low heritability (h^2), suggesting not enough polygenic signals.

Supplementary Notes

Supplementary Note 1: Details for UKB on phenotyping

In **UKB**, BCC or SCC cases were defined as participants with an ICD-9 or ICD-10 diagnosis code for BCC or SCC and based on histology data (e.g. field ID: 40011). Especially, BCC/SCC cases were defined using the following codes:

BCC case definition: have at least one ICD10 or ICD9 diagnosis

#ICD10 invasive

anything starting C44 (e.g. C447, C442 etc) in f.40006.0.0 PLUS 8090-8094 or 8097-98 in f.40011.0.0 or

anything starting C44 (e.g. C447, C442 etc) in f.40006.1.0 PLUS 8090-8094 or 8097-98 in f.40011.1.0 or

anything starting C44 (e.g. C447, C442 etc) in f.40006.2.0 PLUS 8090-8094 or 8097-98 in f.40011.2.0 or

...etc down to x=9 in f.40006.x.0 and f.40011.x.0

#ICD9 invasive

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.0.0 PLUS 8090-8094 or 8097-98 in f.40011.0.0 or

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.1.0 PLUS 8090-8094 or 8097-98 in f.40011.1.0 or

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.2.0 PLUS 8090-8094 or 8097-98 in f.40011.2.0 or

...etc down to x=10 in f.40013.x.0 and f.40011.x.0

SCC case definition: have at least one ICD10 or ICD9 diagnosis

#ICD10 invasive

anything starting C44 (e.g. C447, C442 etc) in f.40006.0.0 PLUS 8070-8076 or 8078 in f.40011.0.0 or

anything starting C44 (e.g. C447, C442 etc) in f.40006.1.0 PLUS 8070-8076 or 8078 in f.40011.1.0 or

anything starting C44 (e.g. C447, C442 etc) in f.40006.2.0 PLUS 8070-8076 or 8078 in f.40011.2.0 or

...etc down to x=9 in f.40006.x.0 and f.40011.x.0

#ICD9 invasive

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.0.0 PLUS 8070-8076 or 8078 in f.40011.0.0 or

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.1.0 PLUS 8070-8076 or 8078 in f.40011.1.0 or

anything starting 173 (e.g. 1730, 1731 etc) in f.40013.2.0 PLUS 8070-8076 or 8078 in f.40011.2.0 or

...etc down to x=10 in f.40013.x.0 and f.40011.x.0

Supplementary Note 2: Details for each cohort on genotyping, imputation, and quality control

Samples for each cohort (GERA, MGB, UK Biobank, and 23andMe Research cohort) were collected with informed consent and were genotyped on a variety of commercial arrays, as detailed before¹⁻⁷.

GERA

The GERA cohort consists of 110,266 adults who consented to the Research Program on Genes, Environment, and Health at Kaiser Permanente Northern California (KPNC). GERA participants' DNA samples were extracted from Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada) at KPNC and genotyped at the Genomics Core Facility of the University of California, San Francisco (UCSF). DNA samples were genotyped at over 665,000 genetic markers on four ethnic-specific Affymetrix Axiom arrays (Affymetrix, Santa Clara, CA, USA) optimized for European, Latino, East Asian, and African American individuals^{1,2}. Genotype quality control (QC) procedures and imputation were conducted on an array-wise basis⁸, after an updated genotyping algorithm with an advanced normalization step specifically for SNPs in batches not recommended or flagged by the outlier plate detector than has previously been done. Subsequently, variants were excluded if: >3 clusters were identified; the number of batches was <38/42 (EUR array), <3/5 (AFR), <3/6 (EAS), or <7/9 (LAT); and the ratio of expected allele frequency variance across packages was <100 (EUR), <50 (AFR), <100 (EAS), <200 (LAT). On the EUR array, variants were additionally excluded if heterozygosity >.52 or <.02, and if an association test between Reagent kit v1.0 and v2.0 had $P < 10^{-4}$. Imputation was done by array, and we additionally removed variants with call rates <90%. Genotypes were then pre-phased with Eagle⁹ v2.3.2, and then imputed with Minimac3¹⁰ v2.1.1, twice, using two reference panels, in a manner almost identical to that done in the UKB⁶, and described previously¹¹. In detail, the two reference panels included (1) the EGA release of the Haplotype Reference Consortium (n=27,165; no indels)¹² and (2) the 1000 Genomes Project Phase 3 (N=2,404; e.g., indels)¹³. The first reference panel (1) uses a larger sample size (including all the individuals in (2)) and should theoretically impute better, but it is only available at non-indels due to harmonization issues; the second reference panel (2) includes, e.g., indels, which are not present in the first panel, but at a smaller sample size. Thus, we imputed using both reference panels, but if a variant was in both (1) and (2), we used that which should have imputed better, ie., (1). As above-mentioned imputation in GERA was almost identical to that of UKB as has been described⁶.

MGB

The MGB Biobank is an extensive integrated database containing clinical data from MGB HealthCare for ~100,000 consented patients and genomic data for over 35,000 participants⁵. MGB samples were genotyped using three versions of SNP array offered by Illumina (Illumina, Inc., San Diego, CA), including 1) Multi-Ethnic Genotyping Array (MEGA) array including 1,416,020 SNPs, 2) Expanded Multi-Ethnic Genotyping Array (MEGA Ex) array including 1,741,376 SNPs, 3) Multi-Ethnic Global (MEG) array including 1,778,953 SNPs, and 4) Global Screening Array (GSA). The Genome Reference Consortium Human genome build 37 (GRCh37) was used in annotating variants. Imputation was performed using the Michigan Imputation Server that uses Minimac3¹⁰. MGB Biobank uses the HRC reference panel (Version r1.1 2016) for imputation. This HRC panel consists of 64,940 haplotypes of predominantly European ancestry. Haplotype phasing was performed using SHAPE-IT¹⁴. A principal

component analysis (PCA) was applied to characterize population structure and exclude racial outliers. QC steps of genotyped data were conducted for the PCA. Briefly, any variants with a SNP call rate <98% and MAF < 0.01, as well as any subjects with sample call rate < 98%, a discrepancy between the reported and predicted sex, evidence of an excess of homozygosity, or related or duplicated subjects (identity-by-descent [IBD] > 0.2) were excluded from the PCA. For the genome-wide association analyses, imputed SNPs were used. Only common variants of four arrays (MEG, MEGA, MEGA EX, GSA-A) were included in further analyses after conducting QC for each array. Autosomal biallelic SNPs with an info score > 0.8 (high-quality imputed SNPs), call rates > 95%, and MAF > 1% were retained for the association analyses.

UK Biobank

The UK Biobank is a large prospective study following the health of approximately 500,000 participants resident in the UK aged between 40 and 69 years-old at the baseline recruitment visit⁷. Genotyping and imputation of UK Biobank samples were carried out by members of the UK Biobank team, and more information on participant recruitment, genotyping and imputation have been published elsewhere^{6,7}. Briefly, individuals were genotyped using the UK BiLEVE Axiom and UKB Axiom Arrays⁶. After genetic quality control procedures, imputation to the two reference panels (1) the full HRC (n=32,488; no indels) and (2) a merger of 1000 Genomes Phase 3 and the UK10K (n=6,285; indels) was conducted as previously described⁶.

23andMe Research Cohort

Genotyping of the 23andMe samples was conducted using four different arrays (V1, V2, V3, V4), that included custom variants and content from the Illumina HumanHap550+ BeadChip, and Illumina OmniExpress + BeadChip arrays. Imputation to version 3 release of the 1000 Genomes reference haplotypes¹⁵ was conducted using Minimac2^{4,16}. The full GWAS descriptions, including details on data collection, genotyping, imputation and quality control) of 23andMe research cohort as a discovery GWAS (for both BCC and SCC) have been previously reported^{3,4}.

Supplementary References

- 1 Hoffmann, T. J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89 (2011). <https://doi.org/10.1016/j.ygeno.2011.04.005>
- 2 Hoffmann, T. J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-430 (2011). <https://doi.org/10.1016/j.ygeno.2011.08.007>
- 3 Chahal, H. S. *et al.* Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. *Nat Commun* **7**, 12048 (2016). <https://doi.org/10.1038/ncomms12048>
- 4 Chahal, H. S. *et al.* Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nat Commun* **7**, 12510 (2016). <https://doi.org/10.1038/ncomms12510>
- 5 Weiss, S. T. & Shin, M. S. Infrastructure for Personalized Medicine at Partners HealthCare. *J Pers Med* **6** (2016). <https://doi.org/10.3390/jpm6010013>
- 6 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018). <https://doi.org/10.1038/s41586-018-0579-z>
- 7 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015). <https://doi.org/10.1371/journal.pmed.1001779>
- 8 Kvale, M. N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051-1060 (2015). <https://doi.org/10.1534/genetics.115.178905>
- 9 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016). <https://doi.org/10.1038/ng.3679>
- 10 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016). <https://doi.org/10.1038/ng.3656>
- 11 Hoffmann, T. J. *et al.* A large genome-wide association study of QT interval length utilizing electronic health records. *Genetics* **222** (2022). <https://doi.org/10.1093/genetics/iyac157>
- 12 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016). <https://doi.org/10.1038/ng.3643>
- 13 Birney, E. & Soranzo, N. Human genomics: The end of the start for population sequencing. *Nature* **526**, 52-53 (2015). <https://doi.org/10.1038/526052a>
- 14 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181 (2011). <https://doi.org/10.1038/nmeth.1785>
- 15 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010). <https://doi.org/10.1038/nature09534>
- 16 Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-784 (2015). <https://doi.org/10.1093/bioinformatics/btu704>