GCPBayes pipeline: a tool for exploring pleiotropy at the gene level

Yazdan Asgari, Pierre Emmanuel Sugier, Taban Baghfalaki, Elise Anne Lucotte, Mojgan Karimi, Mohammed Sedki, Amélie Ngo, B. Liquet, Thérèse Truong

Published online 5 July 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 3 **1**
https://doi.org/10.1093/nargab/lqad065

GCPBayes pipeline: a tool for exploring pleiotropy at the gene level

Yazdan Asgari[1,*,†], Pierre-Emmanuel Sugier[1,2,†], Taban Baghfalaki[3], Elise Lucotte[1], Mojgan Karimi[1], Mohammed Sedki[4], Amélie Ngo[1], Benoit Liquet[2,5] and Thérèse Truong[1]

[1]Paris-Saclay University, UVSQ, Gustave Roussy, Inserm, CESP, Team Exposome and Heredity, 94807 Villejuif, France, [2]Laboratoire de Mathématiques et de leurs Applications de Pau, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, E2S-UPPA, 64000 Pau, France, [3]Inserm U1219, Univ. Bordeaux, ISPED, 33076 Bordeaux, France, [4]Paris-Saclay University, UVSQ, Gustave Roussy, Inserm, CESP, Team Psychiatrie du développement et trajectoires, 94807 Villejuif, France and [5]School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW 2109, Australia

Received January 04, 2023; Revised May 16, 2023; Editorial Decision June 13, 2023; Accepted June 16, 2023

ABSTRACT

Cross-phenotype association using gene-set analysis can help to detect pleiotropic genes and inform about common mechanisms between diseases. Although there are an increasing number of statistical methods for exploring pleiotropy, there is a lack of proper pipelines to apply gene-set analysis in this context and using genome-scale data in a reasonable running time. We designed a user-friendly pipeline to perform cross-phenotype gene-set analysis between two traits using GCPBayes, a method developed by our team. All analyses could be performed automatically by calling for different scripts in a simple way (using a Shiny app, Bash or R script). A Shiny application was also developed to create different plots to visualize outputs from GCPBayes. Finally, a comprehensive and step-by-step tutorial on how to use the pipeline is provided in our group's GitHub page. We illustrated the application on publicly available GWAS (genome-wide association studies) summary statistics data to identify breast cancer and ovarian cancer susceptibility genes. We have shown that the GCPBayes pipeline could extract pleiotropic genes previously mentioned in the literature, while it also provided new pleiotropic genes and regions that are worthwhile for further investigation. We have also provided some recommendations about parameter selection for decreasing computational time of GCPBayes on genome-scale data.

INTRODUCTION

Genome-wide association studies (GWAS) have been successful in identifying hundreds of thousands of single-nucleotide polymorphisms (SNPs) associated with risk of complex traits, and it was shown that the majority are associated with more than one trait, suggesting that pleiotropy (i.e. the fact that one genetic variant can affect multiple traits) is a widespread phenomenon in human diseases (1). Studying pleiotropy could help to identify shared biological mechanisms and disentangle relationships between associated diseases and could lead to new perspectives for prevention strategy and for treatment (2).

Different statistical approaches have been proposed to explore pleiotropy between different traits by testing for cross-phenotype (CP) associations at the SNP level (3). A CP association is defined by a genetic locus associated with more than one trait in a study, regardless of the underlying cause for the observed association. A genetic locus is said pleiotropic when it truly affects more than one trait and is one possible underlying cause for an observed CP association. There exist several methods that were developed to detect CP association across two or more traits for a given SNP, such as ASSET, a subset-based meta-analysis approach that considers a null hypothesis in which no association exists between a given variant and any traits (4). Therefore, a rejection of the null hypothesis means that the variant is associated with at least one trait and complementary procedures are needed to detect variants associated with more than one trait. Recently, a new SNP-level pleiotropy method has been proposed (PLACO) that allows a user to explore pleiotropy at the SNP level under the null hypothesis that a variant is associated with none or only one of the traits (5). Another method (CPBayes) has been proposed to measure the evidence of

*To whom correspondence should be addressed. Tel: +33 1 77 74 15; Email: yazdan.asgari@inserm.fr
†The authors contributed equally to this work.

aggregate-level CP association by using a Bayesian framework based on a spike and slab prior, which is commonly used in solving two-class classification problems and therefore allows the selection of an optimal subset of traits associated with a specific locus (6). However, a problem for these kinds of approaches considering CP association at the SNP level only is that a user needs to perform further functional annotation analysis on the SNPs with potential pleiotropic effects on both traits to find potentially implicated genes and pathways, which is challenging in most cases. Another perspective would be performing an SNP-to-gene annotation at the first place to explore pleiotropy at the gene level to make the results easier to interpret. Furthermore, the structure of the common mechanisms shared by multiple phenotypes can be more complex than SNP-level pleiotropy, as different variants in the same locus can be associated with multiple traits, affecting the same gene, and therefore can have an impact on the same protein. The fact that these complex mechanisms are not fully explored in traditional GWAS approaches has been advanced as one possible explanation of the 'missing heritability' in complex diseases (7). Recently, large-scale cross-trait analyses have shown interest in gene prioritization and enrichment on biological pathways as secondary analyses (8,9). Thus, taking into account the group structure directly in the meta-analysis could help to discover novel genetic variants associated with multiple diseases (10). We developed a Bayesian meta-analysis method (called GCPBayes) that can detect CP association both at the group level (gene or pathway level) and within groups (e.g. at the SNP level) (10).

Although there is an increment in publicly available GWAS summary statistics data, it is not trivial to use such data very easily while working with statistical packages for pleiotropy detection since various inputs including different file formats are needed for each package. Besides, some statistical methods are very time-consuming in the case of working with large GWAS data that included several millions of variants. Therefore, it seems necessary to develop user-friendly platforms that could accelerate exploring pleiotropy analysis between pair of traits using GWAS summary statistics data. Here, we have introduced a new pipeline designed to handle genome-level GWAS summary statistics data as inputs and would eventually perform a feasible CP analysis by using the GCPBayes package. The pipeline checks the integrity of the summary statistics data, harmonizes the referent alleles between the different datasets and provides the gene annotations for each variant. The pipeline also contains some functions and a Shiny application for visualization of results, which give an overall overview to a user for further complementary analyses. To illustrate the application of the pipeline, we used it to identify CP associations at the gene level between breast cancer (BC) and ovarian cancer (OC) using GWAS summary statistics data from two large consortia: the Breast Cancer Association Consortium (BCAC) (11) and the Ovarian Cancer Association Consortium (OCAC) (12). We used GCPBayes and compared the highlighted genes with previous findings from the literature.

MATERIALS AND METHODS

Pipeline overview

There are four main sections available throughout the GCPBayes pipeline. A general overview of the pipeline is shown in Figure 1. As shown, three sections (Standardization, Annotation and Running GCPBayes) run for every pair of traits, while there is also an optional section (Linkage Disequilibrium (LD) Clumping). Additionally, we provide an explanation of the analysis functions for input/output files through a visualization section. In Figure 1, we provided a simplified version, while more details about each section, every procedure, the scripts, and input and output files are provided in Supplementary File S1 and in our GitHub page for advanced users.

There are at least three files needed for running the GCPBayes pipeline: GWAS summary statistics data for two traits and an annotation file that contains information about gene locations in human assembly. We considered an annotation file from the GENCODE project (13) into the GCPBayes pipeline. However, it is possible for a user to replace it with any other annotation file.

In the Standardization section, GWAS summary statistics inputs are checked and harmonized. An important issue a user should consider before exploring pleiotropic effect on two traits is to make sure that the panel of SNPs is the same in both datasets, with valid alleles reported. Also, even it is not required, this is more convenient to analyze final results with similar effect alleles on the same strand in both datasets. If the alleles do not match in the two studies and switching alleles is needed, a user should notice to change the sign of beta values for the corresponding SNPs. Further information and details about running this step are provided in the GitHub page and Supplementary File S1 (Section 2.1).

In the Annotation section, the pipeline annotates SNPs into genes according to their coordinates, in order to perform CP analysis at the gene level. The final created files are in Rdata file formats and compatible with the inputs of the GCPBayes package. When the number of SNPs included in a gene is too large, it might be necessary to shrink the number of SNPs within a gene to reduce the computational time. So, we have designed a section in the GCPBayes pipeline to perform LD clumping (14) to select the most significant SNP within an LD block (with default value for r^2 threshold set to 0.8) based on the CP association estimated by PLACO (5) and LD structure information from a 1000 Genomes Phase 3 reference panel. While other methods, such as MTAG (15), LOGTRAM (16) and coloc (17), are available to identify SNPs with pleiotropic effects, we chose to use PLACO in our pipeline due to its unique ability to build SNP-level *P*-values under the null hypothesis that a variant is associated with none or only one of the traits. We acknowledge that other methods exist and may offer distinct advantages, but the use of PLACO was justified in our study based on its specific features. More information about how to prepare the annotation files for the SNPs with or without the LD clumping option is provided in the GitHub page and Supplementary File S1 (Sections 2.2 and 2.3).

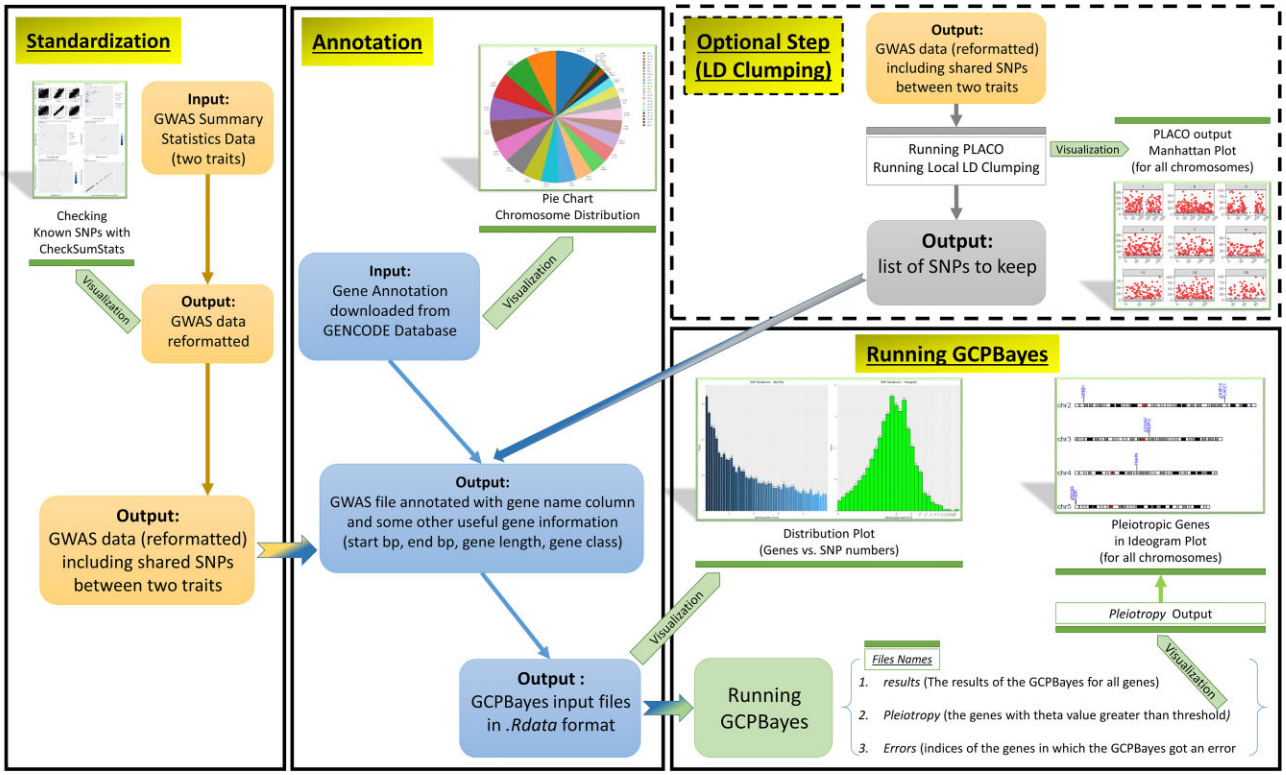

Figure 1. A general overview of the main steps of the GCPBayes pipeline.

In the Running GCPBayes section, the pipeline reads the Rdata files created from the previous section and performs the GCPBayes method to find genes with potential pleiotropic effects on both traits (Supplementary File S1, Sections 2.4.2, 2.4.3, 2.5.2 and 2.5.3). We considered genes with $\theta > 0.5$ as potential pleiotropic genes, as explained in (10), where θ is defined as the probability of having group pleiotropy. So, $\theta > 0.5$ means that the probability to have a pleiotropic effect for the group is $>50\%$. We used this threshold as explanatory analysis in order to generate new candidate genes. However, for a more general use, we implemented a q-value to control for the false discovery rate. We recommend to the user to report on the q-value that we have implemented in the pipeline outputs in order to get more confidence on the results. However, it should be noted that this correction is at least as much conservative as a Benjamini–Hochberg correction.

The current version of the pipeline is running GCPBayes by using the Dirac spike (DS) function (10), but we recommend to then perform analysis on the subset of genes with $\theta > 0.1$ by using the hierarchical spike (HS) function to confirm the results. Then, the same threshold as for DS can be considered. As we explained in more detail in our previous study, DS runs a Gibbs sampler for a multivariate Bayesian sparse group selection model along with DS prior, while HS runs a Gibbs sampler with HS prior to detect a pleiotropic effect (10). Also, HS could perform a more relevant selection of signal at the SNP level as this method is designed for variable selection at both SNP and gene levels. However, analysis with GCPBayes by considering the HS function is more time-consuming than using DS (10).

Due to higher running time of the GCPBayes method while dealing with genes that include a large number of SNPs, the pipeline is also designed to divide genes into two groups according to whether their number of SNPs is lower or higher than a threshold that needs to be defined. It is then possible to restrict the GCPBayes analysis to genes with a number of SNPs lower than the given threshold (Supplementary File S1, Sections 2.4.2 and 2.5.2). Genes with a number of SNPs higher than the given threshold could also be processed through a separate script (Supplementary File S1, Sections 2.4.3 and 2.5.3) or a user could shrink the number of SNPs by considering the LD clumping step. More information is provided in the GitHub page and Supplementary File S1 (Section 2.3).

There are two ways to launch the GCPBayes pipeline: (i) by modifying directly the 'Definition section' that includes all file names and paths related to input and output files, as well as threshold values (if needed) for each section, and by running the Bash or R script that will read all defined parameters and run every step of the pipeline; or (ii) by using the Shiny app we developed that will create the parameter file from the information the user provides (such as working directory path, GWAS column names, threshold values, etc.) and that will run each step of the pipeline sequentially. An R script is also available (which could be run individually or through the Shiny app) to install required packages for the pipeline. More information is provided in the GitHub page.

Visualization analyses

Some visualization parts have been designed throughout the GCPBayes pipeline as follows: checking GWAS summary statistics input data (files created in the Standardization section) by comparing SNP allele effects with previous GWAS results referenced in the GWAS catalog and SNP allele frequencies with those references in 1000 Genomes using a recently developed R package called CheckSumStats (18) (Supplementary File S1, Section 3.1), an overview from the annotation file (Supplementary File S1, Section 3.2), analysis of PLACO outputs created in the LD Clumping section (Supplementary File S1, Section 3.3), analysis of GCPBayes input file created at the end of Annotation section (Supplementary File S1, Section 3.4) and analysis of GCPBayes output file created in the Running GCPBayes section (Supplementary File S1, Section 3.5).

We have also designed a Shiny application to automatically create different graphs based on outputs of the pipeline. The goal of the visualization part is to get an overview by plotting the inputs/outputs and decide about further analyses for a specific region or doing complementary procedures. More details about how to perform the visualization parts are provided in the GitHub page and Supplementary File S1 (Section 3.6).

GWAS data

GWAS summary statistics data for BC risk downloaded from the BCAC (version of year 2020) included 10 760 767 SNPs from 133 384 BC cases and 113 789 controls (11). For OC, data downloaded from the OCAC contained summary statistics for 18 119 090 SNPs derived from 25 509 epithelial OC cases and 40 941 controls in total (12). Some preprocessing steps were performed on the data (as explained in the Standardization section and also in Supplementary File S1, Section 2.1). The reformatted data were used through the GCPBayes pipeline for further analyses.

RESULTS

In the Standardization steps (Figure 1), we removed all SNPs with missing effect or incomplete allele information. For the BCAC, a total of 9 149 691 SNPs remained, while the final OCAC GWAS summary statistics data included 8 929 032 SNPs. The OCAC GWAS data were also recoded in order to keep the same referent and effect alleles as in the BCAC data. For details about cleaning and reformatting steps, see Supplementary File S1 (Section 2.1). By keeping shared entries for BCAC and OCAC GWAS summary statistics data, 7 079 969 SNPs were considered for further analyses. We first compared the GWAS summary statistics data from the two datasets to find out common significant loci. As shown in Supplementary Figure S1, there were three loci (5p15.33, 10p12.31 and 15q26.1) that were significant in both BC and OC data.

After running the pipeline without performing LD clumping, 7 079 969 SNPs were mapped into 18 244 protein-coding genes of which 18 124 genes contained $\leq$1500 SNPs, while 120 genes had >1500 SNPs (Supplementary Table S1). For the second group of the genes with >1500 SNPs, we used the GCPBayes method with the LD clumping step in order to shrink the SNP numbers by prioritizing the SNPs that have the most significant CP associations within an LD block using the PLACO method. After running GCPBayes on two groups of genes, 151 protein-coding genes across 79 different genomic regions were selected as potential pleiotropic genes ($\theta > 0.5$) between BC and OC throughout autosomal chromosomes (Supplementary Table S2). A summary of the gene properties with potential pleiotropic effects distributed by chromosomes is provided in Table 1. GCPBayes detected genes for which the number of SNPs ranges from 1 to 962. All pleiotropic genes were detected through the SNPs that were not clumped.

The reformatting of BCAC and OCAC GWAS data for all chromosomes and the annotation using protein-coding genes took, respectively, ~20 and ~8 min without and with the LD clumping step (on a server with Intel® Xeon® Processor E7-8860 v4 @ 2.20 GHz, 516 GB RAM, CentOS 7.9.2009). It should be noted that the pipeline could be run in a regular PC with smaller size of RAM (e.g. Intel® Core™ i7-1165G7 @ 2.80 GHz, 32 GB RAM) in ~30 min. We have also shown running times for some genes with different numbers of SNPs using the DS function of the GCPBayes package to give to the user an estimation for the package running time related to various numbers of SNPs (Table 2). We observed that the running time increases with a higher number of SNPs. So, we recommend to use the LD clumping step to shrink the number of SNPs for genes with a very high number of SNPs.

While using the LD clumping step, it should be noted that some genes could be removed. Also, while the number of SNPs for a gene is still greater than the chosen threshold after LD clumping, this gene is not included in the GCPBayes analysis. Therefore, selection of a smaller value for the LD clumping threshold makes the pipeline run faster but would cause losing more genes. We performed the analyses based on different threshold values and finally suggest a threshold of 1500 SNPs (Supplementary Table S3 and Supplementary Figure S2). Using this threshold, we performed the LD clumping step for 120 genes out of 18 244 coding genes, and running the whole process took ~40 h using only six CPUs, while we missed two genes (one during the LD clumping step and one due to higher number of SNPs than the LD threshold).

We compared the 79 highlighted loci with findings from previous studies that analyzed the pleiotropic effects between BC and OC (Table 3).

Looking through the literature, we identified eight previous studies that reported only a few overlapping susceptibility loci reported between the two cancers (Supplementary Table S4). Four of these previous studies focused on a specific locus (5p15.33, 8q24.21, 15q26.1 and 19p13.11) (19–22), while three others conducted a genome-wide meta-analysis at the SNP level (23–25) and one study used transcriptome-wide association studies (TWAS) to predict pleiotropic genes in the genome (26). These studies identified 44 potential pleiotropic loci of which 31 were reported by the studies with a GWAS approach. Our study could retrieve 15 of these 31 loci and 18 out of all the 44 previously reported loci as having potential pleiotropic effects on BC and OC (Supplementary Table S4). Our study also suggested 61 new loci with potential pleiotropic signals. We

Table 1. Summary for number of genes in each chromosome with potential pleiotropic signals between BCAC and OCAC GWAS data obtained after running the pipeline (more details are available in Supplementary Table S2)

Chromosome	# Genes	Minimum gene length (bp)	Maximum gene length (bp)	Minimum SNP numbers	Maximum SNP numbers
1	5	4959	57 942	3	98
2	5	49 067	394 420	79	672
3	12	33 195	629 281	18	835
4	10	19 285	887 472	32	962
5	6	16 272	96 657	37	472
6	8	40 931	472 928	55	938
7	2	37 608	261 613	88	278
8	3	97 883	158 920	178	447
9	5	10 883	226 958	18	320
10	10	12 893	354 629	3	446
11	8	11 010	255 035	11	530
12	14	5150	448 262	2	801
15	9	8255	225 330	18	364
16	7	11 508	88 046	12	197
17	22	444	371 278	1	941
19	13	3130	123 106	1	143
20	7	6807	178 396	10	162
21	2	49 232	104 701	82	151
22	3	15 476	701 851	30	652

Table 2. Running time used for calculation of the DS function in the GCPBayes package for some genes with different numbers of SNPs

Gene name	# SNPs	Running time		
		Seconds	Minutes	Hours
ACTL7A	1	2.917		
AP2S1	2	3.169		
ARHGAP5	5	3.239		
CD37	10	3.479		
ARSF	50	8.412		
FAF2	100	24.295		
DDR2	200		3.51	
KIF13B	300		8.46	
TTC27	400		17.24	
PPP3CA	500		30.63	
AMPH	699		77.44	
CDH13	1031			3.95
RBFOX1	1424			9.70
CSMD1	2264			38.77

were also able to retrieve all three loci (Table 3) that were significant through direct comparison of GWAS summary statistics provided in Supplementary Figure S1.

We used the Shiny application to visualize the candidate pleiotropic genes, which reads the GCPBayes pleiotropy output and creates different graphs such as scatter plot for any pair of numeric data from the GCPBayes output, histogram for theta values, distribution of gene lengths and number of SNPs as a circus plot, distribution of pleiotropic genes on chromosomes as a pie chart, gene lengths as box and violin plots, and gene locations on chromosomes as a karyotype plot (for each chromosome separately and for all chromosomes in one graph). The graphs generated for BCAC and OCAC data are provided in Supplementary Figure S3.

DISCUSSION

We have provided a pipeline for our recently developed method GCPBayes (10) that explores pleiotropy at the gene level. The pipeline starts with standardization of GWAS summary statistics data for two traits and performs an annotation step, creates inputs for the GCPBayes package and then provides results for CP association at the gene level using GCPBayes. We have also developed a visualization tool using Shiny application that loads the GCPBayes output data and creates various plots automatically.

To illustrate applicability of the pipeline, we run it on GWAS data for BC and OC. A total of 151 genes from 79 loci were detected to be associated with BC and OC (Supplementary Table S2).

We could retrieve three genes (*TERT*, *RCCD1* and *BABAM1*) that have been reported to be associated with both BC and OC in previous fine-mapping studies at 5p15.33, 15q26.1 and 19p13 loci (19,21,22). We also highlighted the 8q24.21 locus that was reported to be associated with multiple cancers (20).

Fehringer *et al.* (25) have previously explored pleiotropy across multiple cancers (lung, ovarian, breast, prostate, colorectal) using the method ASSET that conducts a fixed effect meta-analysis by examining the association between each SNP and multiple subsets of cancer. This study reported 130 SNPs from 21 loci to be pleiotropic between BC and OC (Supplementary Table S4). Our study could replicate seven loci (3p24.1, 5p15.33, 5q11.2, 9p21.3, 10q26.13, 19p13.11 and 20q11.22) reported in their study (Supplementary Table S4).

In (23), an SNP-level cross-cancer genome-wide association meta-analysis focusing on breast, ovarian and prostate cancers was conducted using GWAS statistics data from the BCAC and OCAC based on a lower number of individuals than the datasets we used. They reported seven loci associated with both BC and OC risks. The same authors updated in 2020 (24) their results with the most recent GWAS summary statistics data to analyze CP association between breast, prostate, ovarian and endometrium cancers. They reported four new loci with shared association with BC and OC risks. In the current study, we replicated 5 of the 11 suggested pleiotropic loci provided in these two studies.

Table 3. List of loci with potential pleiotropic signals (results of the GCPBayes pipeline) and comparison with reported loci in the literature

Chromosome	Loci	Known gene	References
1	1p34.1, 1p13.2, 1q21.1, 1q32.1		
2	2p25.1, 2p23.2, 2p24.3, 2q13, 2q33.1		
3	3p25.3, 3p24.1, 3p11.1, 3q12.1, 3p13, 3q25.31		
4	4p14, 4q13.1, 4q21.22, 4q31.1, 4q31.21, 4q34.1		
5	5p15.33	*TERT* ($\theta = 1$)	(21)
	5q11.2, 5q31.1		
6	6p23, 6q22.31, 6q22.33, 6q24.3, 6q25.1, 6q26		
7	7q21.3, 7q22.1		
8	8q24.21, 8p11.23, 8q21.13		
9	9q31.1	*SMC2* ($\theta = 1$)	(23)
	9p21.3, 9q21.13, 9q33.2, 9q34.2		
10	10p15.1, 10p12.31, 10p11.22, 10q21.3, 10q24.32, 10q25.2, 10q25.3, 10q26.13		
11	11p15.5, 11p11.2, 11q13.2, 11q13.3, 11q23.3		
12	12p11.22, 12q13.2, 12q15, 12q24.11, 12q24.13, 12q24.31		
15	15q26.1	*RCCD1* ($\theta = 1$)	(19,23)
	15q15.1, 15q22.31		
16	16p12.2, 16q22.1, 16q23.2		
17	17p12, 17p13.1, 17q21.2, 17q21.31, 17q21.32, 17q22, 17q25.1		
19	19p13.11	*BABAM1* ($\theta = 0.986$)	(22)
	19p13.2		
20	20q11.22	*CPNE1* ($\theta = 0.847$)	(26)
	20q13.33	*RGS19* ($\theta = 1$)	(26)
21	21q22.12	*CLIC6* ($\theta = 1$)	(24)
	21q21.1		
22	22q12.1		

Besides, in a TWAS study, Kar *et al.* reported 14 loci with shared association with BC and OC (26). Our method could retrieve four loci (7q21.3, 16q22.1, 20q11.22 and 20q13.33) in common with their study (Supplementary Table S4).

In summary, previous studies identified a total of 44 loci with potential pleiotropic effects on BC and OC. Using GCPBayes, we found 79 loci in our study of which 18 loci were reported in the previous studies. Further analyses would be needed to confirm the potential pleiotropic effect between BC and OC of the remaining 61 loci.

However, as every study, our approach contains some limitations. GCPBayes could not recapture all previously published genes with pleiotropic signals. One example was 13q13.1 locus, which contains *BRCA2* gene ($\theta = 0$) (25). This locus was not highlighted in our study. In another study, Ghoussaini *et al*. demonstrated *c-MYC* gene (located at 8q24.21 region) with a potential pleiotropic effect between BC and OC based on a fine-mapping approach (20), while our study detected another gene in the same locus (*POU5F1B*) with a potential pleiotropic signal (Supplementary Table S2). This suggests that it might need to use other integrative data such as transcriptome data used by Kar *et al.* (26) as well as in fine-mapping studies in order to improve selection criteria for pleiotropic genes and loci. Indeed, mapping trait-associated SNPs to their nearest gene can fail to identify the functional gene (27). Several methods have been developed to improve the functional relevance of SNP-to-gene annotations (28) by considering gene expression information based on studies on expression quantitative trait loci (eQTL). It also needs some useful functions such as functional enrichment analysis based on a list of genes selected by using GCPBayes.

There are also some other limitations while working with GWAS summary statistics data that lead to a loss of information about the covariance matrix of beta estimates and then would lead to an underconsideration of the LD structure of the group. GCPBayes uses the diagonal matrix with variance of beta estimates as approximation of the previous matrix that can lead to a loss of power of detection of genes with pleiotropic effects. Further work is ongoing to allow the user to exploit the covariance matrix of the genotypes from a population of reference (1000 Genomes) to approximate the covariance matrix of beta estimates. This approximation has already been exploited by some methods in the field (29).

Another issue when working with the GCPBayes method is that the running time increases significantly with a higher number of SNPs (especially when it is >500). In this study, we proposed to shrink genes with the highest number of SNPs using an LD clumping step. Another possible solution would be to separate these longest genes into multiple subgroups according to their LD block structures.

In conclusion, we proposed an exploratory method that permits to prioritize potential pleiotropic loci and genes for further investigation using GWAS summary statistics data. The proposed pipeline is publicly available at https://github.com/CESP-ExpHer/GCPBayes-Pipeline. In order to resolve the limitations and improve, we will keep updating the pipelines regularly.

DATA AVAILABILITY

All scripts, Bash files and detailed information about how to run on any dataset (also how to run a Bash file on BCAC and OCAC GWAS summary statistics data), and a step-by-step tutorial for running the GCPBayes pipeline on BCAC and OCAC GWAS summary statistics data are available on our group's GitHub page (https://github.com/CESP-ExpHer/GCPBayes-Pipeline, permanent doi: https://doi.org/10.5281/zenodo.8042138). Input GWAS summary statistics data are accessible through the web page of the

BCAC (version 2020) (https://bcac.ccge.medschl.cam.ac.uk/) and OCAC (https://ocac.ccge.medschl.cam.ac.uk/). In addition, information about how to run Bash file on BCAC and OCAC GWAS summary statistics data, and a step-by-step tutorial for obtaining the results using these GWAS input data including all sections are provided in Supplementary File S1. Besides, a comprehensive wiki (manual) for all scripts used in the GCPBayes pipeline is available on our group's GitHub page that could be used by developers to modify/add features to the pipeline.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors acknowledge the BCAC and OCAC, international multidisciplinary consortiums, for providing GWAS summary statistics data. The breast cancer GWAS for the BCAC was supported by Cancer Research UK (PPRPGM-Nov20\100002, C1287/A10118, C1287/A16563, C1287/A10710, C12292/A20861, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565) and the Gray Foundation, the National Institutes of Health (CA128978, X01HG007492—the DRIVE consortium), the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, the Quebec Breast Cancer Foundation, the European Community's Seventh Framework Programme under grant agreement no. 223175 (HEALTH-F2-2009-223175) (COGS), the European Union's Horizon 2020 Research and Innovation Programme (634935 and 633784), the PostCancer GWAS initiative (U19 CA148537, CA148065 and CA148112—the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer (CRN-87521), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund. All studies and funders are listed in (11).

Authors' contributions: Y.A., P.-E.S., T.T., and B.L. designed the study. Y.A., P.-E.S., T.B., E.L., M.S., M.K. and A.N. prepared the data and wrote the codes in the pipeline. Y.A. and P.-E.S. analyzed the data. Y.A., P.-E.S. and T.T. interpreted the application data. Y.A., P.-E.S., B.L. and T.T. were major contributors to writing the manuscript. All authors read and approved the final manuscript.

FUNDING

Ligue Contre le Cancer; Inserm Cross-Cutting Project GOLD; Inserm Itmo Cancer.

Conflict of interest statement. The authors have no conflicts of interest to declare.

REFERENCES

1. Watanabe,K., Stringer,S., Frei,O., Umićević Mirkov,M., de Leeuw,C., Polderman,T.J.C., van der Sluis,S., Andreassen,O.A., Neale,B.M. and Posthuma,D. (2019) A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, **51**, 1339–1348.
2. Gratten,J. and Visscher,P.M. (2016) Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.*, **8**, 78.
3. Solovieff, N., Cotsapas,C., Lee,P.H., Purcell,S.M. and Smoller,J.W. (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.
4. Bhattacharjee,S., Rajaraman,P., Jacobs,K.B., Wheeler,W.A., Melin,B.S., Hartge,P., Yeager,M., Chung,C.C., Chanock,S.J. and Chatterjee,N. (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.*, **90**, 821–835.
5. Ray,D. and Chatterjee,N. (2020) A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *PLoS Genet.*, **16**, e1009218.
6. Majumdar,A., Haldar,T., Bhattacharya,S. and Witte,J.S. (2018) An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS Genet.*, **14**, e1007139.
7. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
8. Zhu,Z., Lee,P.H., Chaffin,M.D., Chung,W., Loh,P.R., Lu,Q., Christiani,D.C. and Liang,L. (2018) A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.*, **50**, 857–864.
9. Márquez,A., Kerick,M., Zhernakova,A., Gutierrez-Achury,J., Chen,W.M., Onengut-Gumuscu,S., González-Álvaro,I., Rodriguez-Rodriguez,L., Rios-Fernández,R., González-Gay,M.A. *et al.* (2018) Meta-analysis of Immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. *Genome Med.*, **10**, 97.
10. Baghfalaki,T., Sugier,P.E., Truong,T., Pettitt,A.N., Mengersen,K. and Liquet,B. (2021) Bayesian meta-analysis models for cross cancer genomic investigation of pleiotropic effects using group structure. *Stat. Med.*, **40**, 1498–1518.
11. Zhang,H., Ahearn,T.U., Lecarpentier,J., Barnes,D., Beesley,J., Qi,G., Jiang,X., O'Mara,T.A., Zhao,N., Bolla,M.K. *et al.* (2020) Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.*, **52**, 572–581.
12. Phelan,C.M., Kuchenbaecker,K.B., Tyrer,J.P., Kar,S.P., Lawrenson,K., Winham,S.J., Dennis,J., Pirie,A., Riggan,M.J., Chornokur,G. *et al.* (2017) Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.*, **49**, 680–691.
13. Frankish,A., Diekhans,M., Ferreira,A.M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
14. Price,A.L., Weale,M.E., Patterson,N., Myers,S.R., Need,A.C., Shianna,K.V., Ge,D., Rotter,J.I., Torres,E., Taylor,K.D.D. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132.
15. Turley,P., Walters,R.K., Maghzian,O., Okbay,A., Lee,J.J., Fontana,M.A., Nguyen-Viet,T.A., Wedow,R., Zacher,M., Furlotte,N.A. *et al.* (2018) Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.*, **50**, 229–237.
16. Xiao,J., Cai,M., Yu,X., Hu,X., Chen,G., Wan,X. and Yang,C. (2022) Leveraging the local genetic structure for trans-ancestry association mapping. *Am. J. Hum. Genet.*, **109**, 1317–1337.
17. Giambartolomei,C., Vukcevic,D., Schadt,E.E., Franke,L., Hingorani,A.D., Wallace,C. and Plagnol,V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.
18. Haycock,P.C., Carolina Borges,M., Burrows,K., Lemaitre,R.N., Harrison,S., Burgess,S., Chang,X., Westra,J., Khankari,N.K.,

Tsilidis,K. *et al.* (2023) Design and quality control of large-scale two-sample Mendelian randomization studies. *International Journal of Epidemiology*, dyad018.
19. Plummer,J., Dezem,F.S., Chen,S.S., Dhungana,S., Wali,D., Davis,B., Kanska,J., Safi,N., Seo,J.-H., Corona,R.I. *et al.* (2020) Transcriptome and interactome analyses identify the TP53 interacting gene RCCD1 as a candidate susceptibility gene at the 15p26.1 breast and ovarian cancer risk locus. bioRxiv doi: https://doi.org/10.1101/2020.09.29.319699, 30 September 2020, preprint: not peer reviewed.
20. Ghoussaini,M., Song,H., Koessler,T., Al Olama,A.A., Kote-Jarai,Z., Driver,K.E., Pooley,K.A., Ramus,S.J., Kjaer,S.K., Hogdall,E. *et al.* (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.*, **100**, 962–966.
21. Bojesen,S.E., Pooley,K.A., Johnatty,S.E., Beesley,J., Michailidou,K., Tyrer,J.P., Edwards,S.L., Pickett,H.A., Shen,H.C., Smart,C.E. *et al.* (2013) Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.*, **45**, 371–384.
22. Lawrenson,K., Kar,S., McCue,K., Kuchenbaeker,K., Michailidou,K., Tyrer,J., Beesley,J., Ramus,S.J., Li,Q., Delgado,M.K. *et al.* (2016) Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. *Nat. Commun.*, **7**, 12675.
23. Kar,S.P., Beesley,J., Al Olama,A.A., Michailidou,K., Tyrer,J., Kote-Jarai,Z.S., Lawrenson,K., Lindstrom,S., Ramus,S.J., Thompson,D.J. *et al.* (2016) Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. *Cancer Discov.*, **6**, 1052–1067.
24. Kar,S.P., Lindström,S., Hung,R.J., Lawrenson,K., Schmidt,M.K., O'Mara,T.A., Glubb,D.M., Tyrer,J.P., Schildkraut,J.M., Chang-Claude,J. *et al.* (2020) Combining genome-wide studies of breast, prostate, ovarian and endometrial cancers maps cross-cancer susceptibility loci and identifies new genetic associations. bioRxiv doi: https://doi.org/10.1101/2020.06.16.146803, 19 June 2020, preprint: not peer reviewed.
25. Fehringer,G., Kraft,P., Pharoah,P.D., Eeles,R.A., Chatterjee,N., Schumacher,F.R., Schildkraut,J.M., Lindström,S., Brennan,P., Bickeböller,H. *et al.* (2016) Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. *Cancer Res.*, **76**, 5103–5114.
26. Kar,S.P., Considine,D.P.C., Tyrer,J.P., Plummer,J.T., Chen,S., Dezem,F.S., Barbeira,A.N., Rajagopal,P.S., Rosenow,W.T., Moreno,F. *et al.* (2021) Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer. *Hum. Genet. Genomics Adv.*, **2**, 100042.
27. Smemo,S., Tena,J.J., Kim,K.H., Gamazon,E.R., Sakabe,N.J., Gómez-Marín,C., Aneas,I., Credidio,F.L., Sobreira,D.R., Wasserman,N.F. *et al.* (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, **507**, 371–375.
28. Gerring,Z.F., Mina-Vargas,A., Gamazon,E.R. and Derks,E.M. (2021) E-MAGMA: an eQTL-informed method to identify risk genes using genome-wide association study summary statistics. *Bioinformatics*, **37**, 2245–2249.
29. Mishra,A. and Macgregor,S. (2015) VEGAS2: software for more flexible gene-based testing. *Twin Res. Hum. Genet.*, **18**, 86–91.