PheGWAS - ADDING A NEW DIMENSION TO ASSOCIATION STUDIES

Gittu George[1], Sushrima Gan[1], Philip Appleby[1], Yu Huang[1], Alex Doney [1], For Inspired investigators

[1]School of Medicine, University of Dundee, Dundee, United Kingdom

Introduction

The potential of personalized medicine has evolved extensively in the last decade with the development of genome-wide association studies (GWAS), which is a powerful method for exploring the genetic architecture underlying diseases and traits affecting humans. These GWAS data visualizations are for “one phenotype many variant” situation. However, when a researcher is interested in pleiotropy and requires to assess if particular variants are associated with a group of phenotypes PheWAS is undertaken which is the inverse of GWAS considers the “one variant many phenotypes”.

We describe an extension of these approaches in a “many genotypes many phenotype” scenario (Fig 1) to visualize comprehensive genome-wide data with phenome-wide data in three-dimensional space. This approach which we refer to as PheGWAS might assist understanding or exploring pleiotropy at scale. This demands for streamlined, systematic and structured data visualization tools which will be easy to understand and will efficiently handle very large volumes of data.

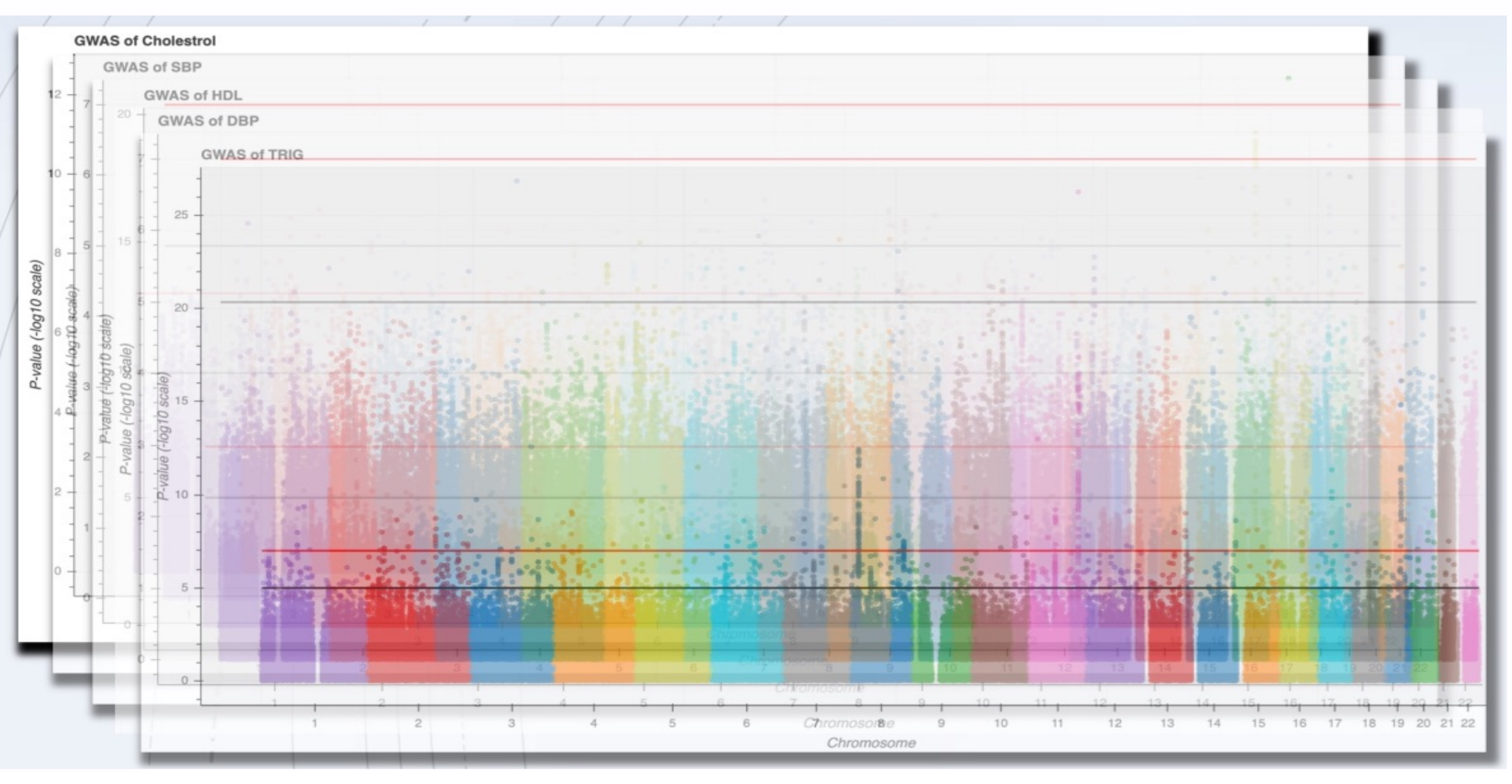

Fig 1: Visualization of foundational backbone of the PheGWAS concept

Conceptualizing PheGWAS

To illustrate, we are using Global Lipids Genetics Consortium (GLGC) summary data. To carry out the analysis, we pass the summary statistics file from GLGC to PheGWAS where it provides informative and interactive visualization at various levels.

- Entire genome level

PheGWAS produces a three-dimensional Interactive landscape visualization (Fig 2). Here the x-axis represents the autosomal chromosomes, the y-axis represents the -log10(P-value) of the GWAS summary statistics and the z-axis represents the phenotypes. The most significant p-value is being selected in each of the respective chromosomes for each phenotype, showcased by the peaks in the graph. The entire genome analysis creates an opportunity for the researchers to select a particular chromosome for further analysis.

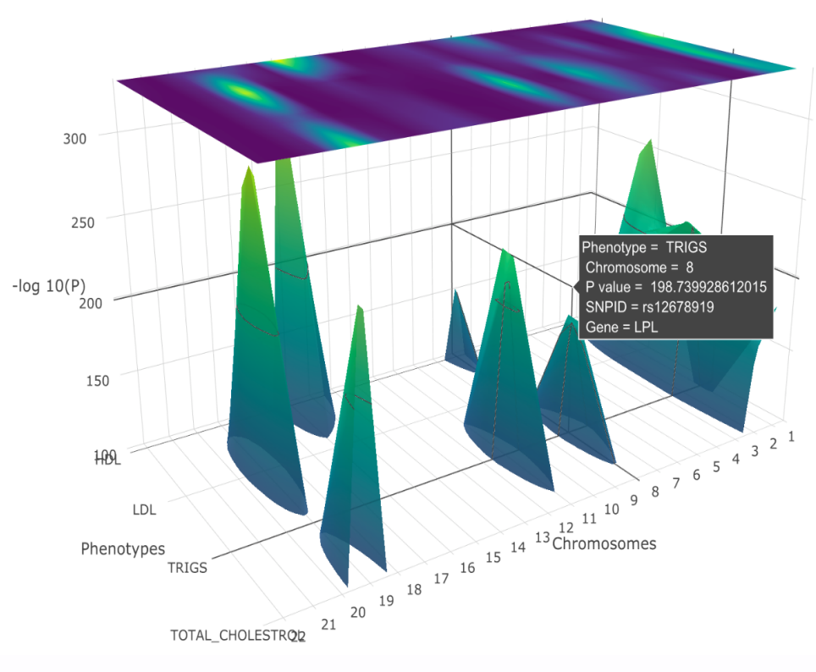

Fig 2: A PheGWAS graph for the phenotypes, SBP, DBP, HDL, Triglycerides and Cholesterol with a sectional view of -log10(p-value) greater than 6.5

Here in Fig 2 in chromosome position 19, we identify two peaks corresponding to LDL and Total Cholesterol. To locate the exact position of the SNPs and to determine pleiotropy , in the next step we select the 19th chromosome for the single chromosomal view.

- Single Chromosomal plot

When a particular chromosome is selected, the entire length of it is divided into user customized segments of base pairs (default 100K base pairs), giving rise to a systematic order of columnar groups. All the peaks are selected which satisfies the underlying condition (Max {–log10(P-Value) >=6}), and this been checked within each phenotype. Here (Fig 3) we get a view of the significant peaks for each of the base pair groups in chromosome 19. We hover the cursor over the heatmap and carry out a comparative analysis on the basis of base pair positions between PheGWAS and GLGC data. Figure 4 shows the heatmaps portraying the identification of different phenotypes at various base pair positions on chromosome 19.

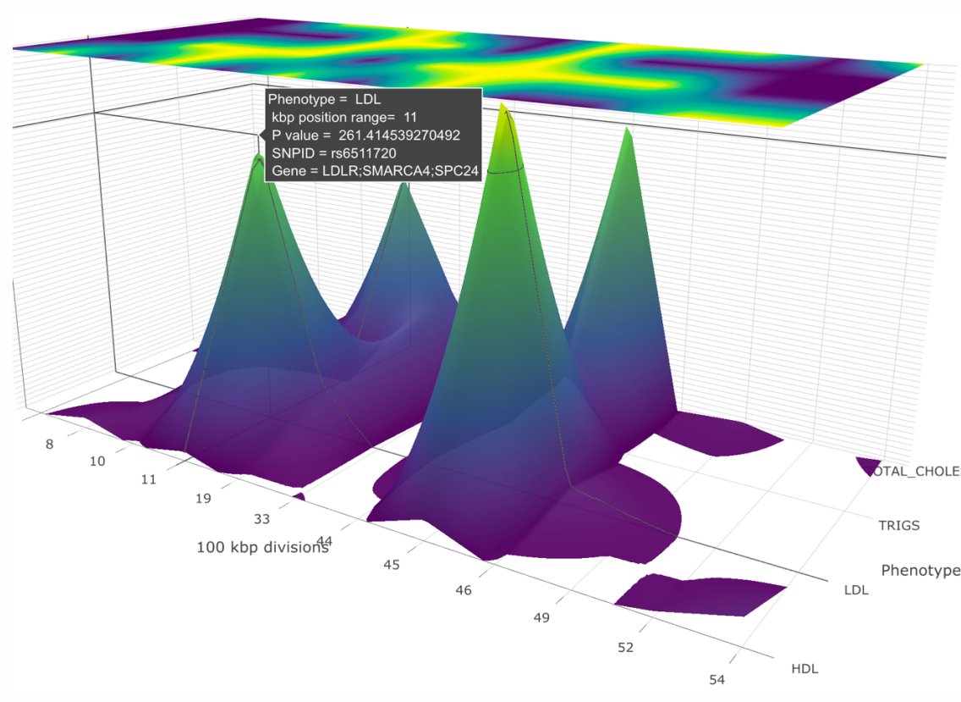

Fig 3: A PheGWAS plot for the sectional view of a single chromosome, produced by plotting the SNPs above a certain threshold of significant values of phenotypes, SBP, DBP, HDL, Triglycerides and Cholesterol

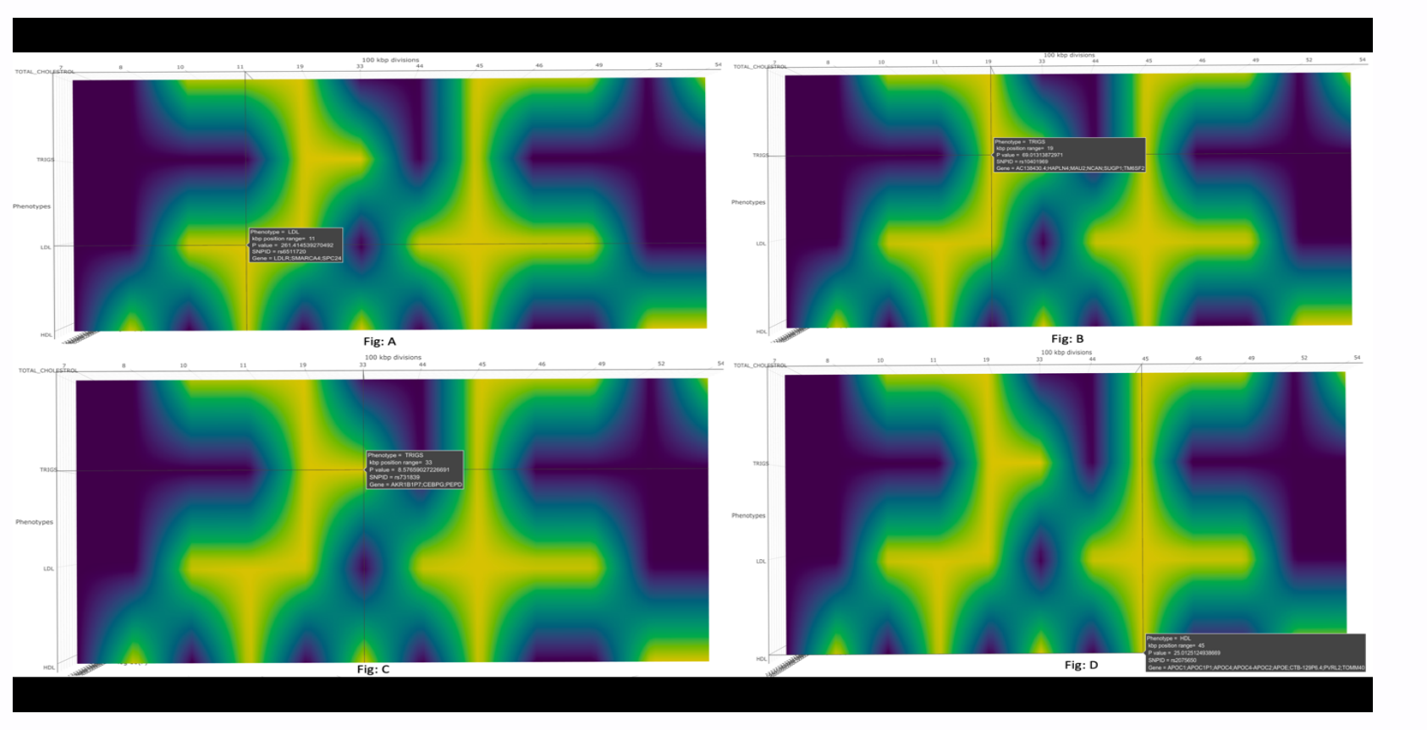

Fig 4: An illustration of the heatmap produced by PheGWAS. The highlighted regions represent the SNPs with significant P-values. This helps the user to decide which all chromosomes will be selected for the individual level chromosomal view.

Implementation

The PheGWAS code has been scripted in R and is wrapped as a package.. The package accepts the GWAS summary files as R data frames to generate the interactive PheGWAS plot. By default, PheGWAS generates an interactive 3D plot for all chromosomes. To plot an individual chromosome, the chromosome number must first be provided as a parameter.

Internally it also feature’s the capability of gene mapping, where SNP ID and genome location data within the GWAS file is mapped to the respective GENE region that it falls into. This mapping is made possible by using the *BioMart* R package, which gives the associated gene for each SNP id in the GWAS summary file.

Validation

For the purpose of validating the PheGWAS method, we have extended the exploration from the 19th Chromosome (Table 1) to all the chromosomes. The genes and SNPs were classified into three groups

i. complete match (CM); if the gene or SNP in a particular PheGWAS group is identified by PheGWAS in all the same traits as identified in the GLGC
ii. partial match (PM); if the gene or SNP is identified in some of the associated traits and not all (applicable only while verifying multiple traits)
iii. no match (NM); if the gene or SNP is not identified by PheGWAS at all.

Nearest Gene	SNP ID	Chromosome	Position MB	Trait	MAF	Allele	Joint p-value	In PheGWAS	Gene Match	SNP Match
1a. Biological Candidate Genes at Novel locus discovered by GLGC and PheGWAS findings										
PEPD	rs731839	19	33.9	TG, HDL	0.35	G/A	3 x 10–9, 3 x 10–9	CM	CM	CM
1b. Joint Meta-analysis association. results. by GLGC and PheGWAS findings										
LDLR	rs6511720	19	11.2000	LDL,TC	.12	T/G	4x10-262/5x10-202	CM	CM	CM
CILP2	rs10401969	19	19.4100	TC,TG,LDL	.09	C/T	4x10-77/1x10-69/3x10-54	CM	NM	CM
APOE	rs4420638	19	45.4200	LDL,TC,HDL	.19	G/A	2x10-178/1x10-149/2x10-21	CM	CM	NM

Table 1: Detailed description of findings of GLGC on Chromosome 19 along with the match status of PheGWAS gene and SNPs

In the entire genome, GLGC identified 157 SNPs ($P<5*10^{-8}$) associated with blood lipid levels (Total Cholesterol, HDL, LDL, Triglycerides), which included 62 newly identified SNPs and 95 previously discovered lipid SNPs. By using PheGWAS technique to display variants associated with various traits from GLGC, we were able to detect all the 157 regions of significance, 125 genes (79.6 %) and 84 SNP’s (53.5 %).

Future Work

We are progressing towards the level succeeding the single chromosomal view, which aims towards the making of a three-dimensional regional scatter for the specific base pair groups. Plans are also there to implement this in Rshiny, so that the user have the flexibility for interacting when giving the parameters

Furthermore, we plan to use PheGWAS in the analysis of retinal traits and drug response using real life data from VAMPIRE and GoDarts datasets, respectively.

Conclusion & Discussion

We have created a three-dimensional plot to maximize the visualization GWAS studies over multiple phenotypes. PheGWAS creates a new visualization approach for “many SNPS against many phenotypes” to aid scientific research

Unlike the static and non-interactive Manhattan plot, PheGWAS does not plot all existing points but only focuses on SNPs over a certain pre-decided level of significance and also provides with the interactive feature

Acknowledgment

All the Post Docs and students in INSPIRED, who gave enough support at various stages of my research .

The research was commissioned by the National Institute for Health Research using Official Development Assistance (ODA) funding [INSPIRED 16/136/102].

Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.