Complex-Traits Genetics Virtual Lab: A community-driven web platform for post-GWAS analyses

Gabriel Cuellar-Partida[1*], Mischa Lundberg[1], Pik Fang Kho[2], Shannon D'Urso[1], Luis F. Gutierrez-Mondragon[3], Liang-Dar Hwang[1]

1. The University of Queensland Diamantina Institute, Brisbane, Queensland, Australia

2. QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

3. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

*Corresponding Author

Abstract

Genome-wide association studies (GWAS) are an important method for mapping genetic variation underlying complex traits and diseases. Tools to visualize, annotate, analyse and query results from these studies can be used to generate hypotheses about the molecular mechanisms underlying the associations. Here, we present the Complex-Traits Genetics Virtual Lab (CTG-VL). In this first release, the CTG-VL integrates over a thousand publicly available GWAS summary statistics, a suite of analysis tools, visualization functions and diverse data sets for genomic annotations. This current release of the CTG-VL also makes available results from pathway- and gene- based analyses from over 1500 complex traits. The CTG-VL complements the landscape of available web applications to yield further insights from these data. In this manuscript, we describe the platform by carrying out various analyses including genetic correlation, Mendelian randomization, pathway- and gene- based analyses using GWAS summary statistics of mood swings derived from UK Biobank.

Availability: The CTG-VL is freely available at https://genoma.io as an online web application.

Contact: g.cuellarpartida@uq.edu.au

Motivation

Genome-wide association studies (GWAS) have revolutionized the genetics research of complex traits and diseases over the last decade [1, 2]. As these studies become more common thanks to initiatives such as UK Biobank [3] and international genetics research consortia, it is increasingly necessary to build user-friendly computational tools to allow researchers without strong expertise in the area to annotate, analyse and share results originating from such studies.

To this end, we present the first release of the Complex-Traits Genetics Virtual Lab (CTG-VL), a web platform to annotate, analyse and share GWAS and post-GWAS results. The CTG-VL complements already widely used web applications including LD-Hub [4], FUMA-GWAS [5] and MR-Base [6] that incorporate some of the functions provided in the CTG-VL. In this manuscript, we introduce the first public release of the CTG-VL, compare it to other platforms and apply some of its functions to GWAS summary statistics of mood swing derived from UK Biobank.

Description

Visualization and annotations

The current release of the CTG-VL facilitates the visual inspection of GWAS results through LocusTrack (Figure 1a) [7] and Manhattan plots (Figure 1b). These visualizations are fully interactive and allow users to annotate and query results with a single click. Amongst current annotations, epigenetic data from ENCODE [8] and Roadmap epigenomics projects [9], eQTL data from the GTEx project [10], GWAS catalogue [11] and polyphen [12]are included. An example of an annotated LocusTrack plot is shown in Figure 1a. Utilities such as obtaining the closest genes for a given genetic variant or CpG, obtaining gene information, SNPs in linkage disequilibrium (LD) and population frequencies are also actionable through a single click. The CTG-VL also provides other graphing/plotting functions, including heatmaps, karyograms and network visualizations (Figure 1c). For the network-visualization, we have incorporated functions to analyse the topology of co-expression networks [13], in addition to others that we consider informative including density, degree distribution, mean degree, and number of edges of the giant component. We have also included routines to estimate shortest paths between nodes.

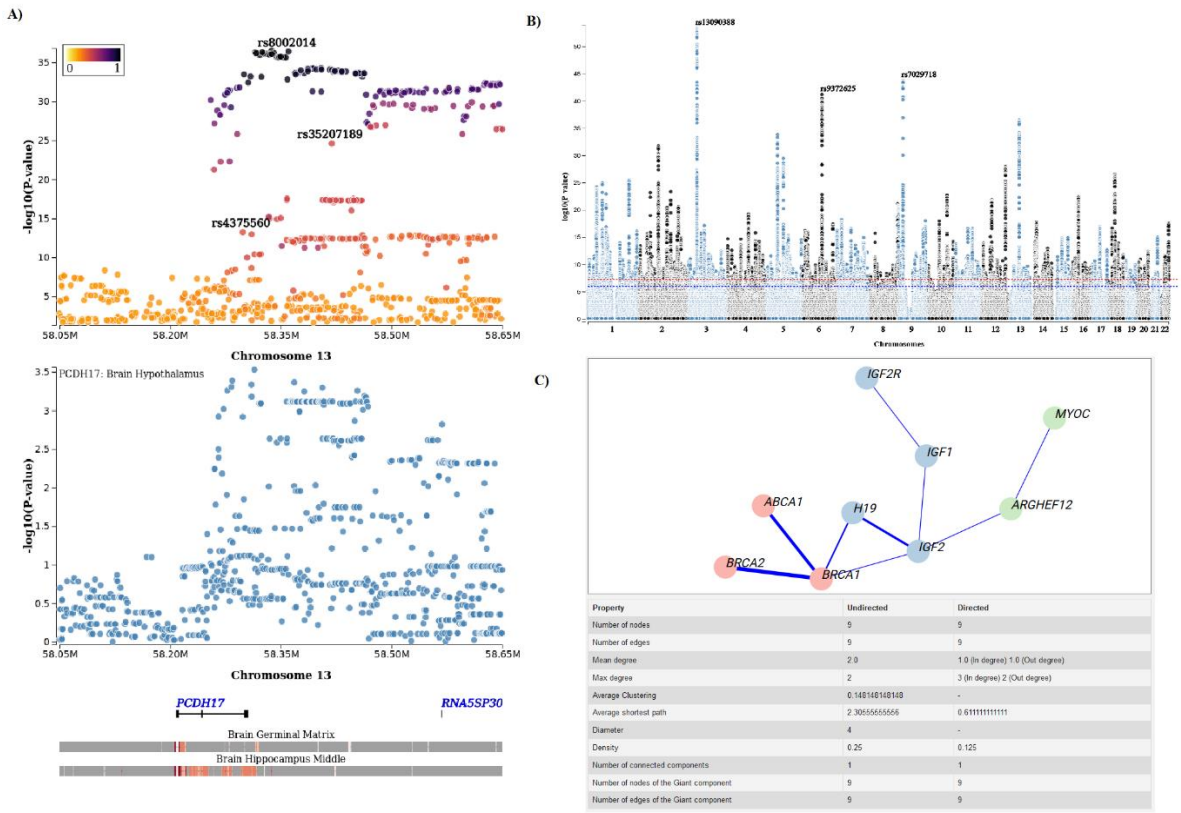

Figure 1. Examples of the visualizations available in the CTG-VL. **(A)** An example of a LocusTrack plot that displays association results within a specified region (upper panel), along with eQTLs for a selected gene (PCDH17 in this example) (middle panel) and annotation tracks such as chromatin information for selected tissues (lower panel). Both, **(A)** LocusTrack and **(B)** Manhattan plots are fully interactive, allowing the user to annotate and query SNPs by simply clicking on them. **(C)** An example of a network visualization and its properties.

Analysis tools

The CTG-VL aims to facilitate the downstream analysis of GWAS summary statistics. As such, we are actively integrating tools and data that will allow the research community to speed up their analysis. The current release of the platform (0.23-alpha) includes some of the newest and most commonly used analysis tools for downstream analysis of GWAS results. Briefly, we have integrated LD-score regression [14, 15] to estimate the heritability of traits and genetic correlation between traits using GWAS summary statistics. Additionally, the user can check genetic correlation against any of the (over 1000) traits with publicly available GWAS summary statistics integrated in the platform. The platform also implements Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT) [16] that prioritizes the most likely causal genes at associated loci, and identifies enriched pathways and tissues/cell types that may underlie the associations. Another popular tool incorporated in CTG-VL is

MetaXcan [17] which tests if the predicted expression levels (based on predicting models derived from eQTL data of selected tissue) of particular genes underlie the association. Finally, we have included the Summary-data-based Mendelian Randomization (SMR) [18] to test the causal relationship between the expression level of a gene and the trait of interest, and the Generalized Summary-data-based Mendelian Randomization (GSMR) [19] to test the causal relationship between two traits.

Data

The CTG-VL aims to function as an aggregator of GWAS summary statistics and post-GWAS analyses results. As such, it complements the recent release of GWAS-ATLAS [20] (http://atlas.ctglab.nl/traitDB/1), a titanic effort that integrates post-GWAS analyses (gene-based and gene-enrichment analyses results from MAGMA [21] as well as genetic correlations and heritability derived from LD-score and SumHer [22]) on over 3500 complex traits in a browsable platform. The CTG-VL has made available post-GWAS results (DEPICT and MetaXcan) of over 1500 traits, and we are actively updating this. In contrast to GWAS-ATLAS, the CTG-VL allows users to link the publicly available GWAS summary statistics in the platform to their user profile where they can perform further analyses.

Web applications ecosystem

The comparison of current web platforms is difficult as each facilitates specific tasks and are optimized for particular pipelines/analysis methods. Given the current ecosystem, we believe that the CTG-VL is a great complement to what is currently available. In Table 1, we briefly describe the differences and similarities between CTG-VL and some of widely used web tools (i.e. FUMA-GWAS, LD-Hub and MR-Base). While each of these existing web tools are aggregators of GWAS summary statistics, their functions are largely segregated. In contrast, the CTG-VL integrates key functions from each platform and others (Table 1).

Table 1. Some of the differences and similarities between CTG-VL, FUMA-GWAS, LD-HUB and MR-Base.

	CTG-VL	**FUMA-GWAS**	**LD-Hub**	**MR-Base**
Aggregator of GWAS summary statistics	Yes	Yes	Yes	Yes
Visualization	All visualizations are interactive: Manhattan plot Regional plot Heatmaps Network plots Karyograms	Manhattan plot Regional plot (interactive) Heatmap (only for gene expression)	No	No
Data for annotation	ENCODE Roadmap Epigenomics GTEx Polyphen	ENCODE Roadmap Epigenomics GTEx Polyphen Multiple others (http://fuma.ctglab.nl/links)	No	No
Heritability	LD-score regression	No	LD-score regression	No
Genetic correlations	LD-score regression The users choose traits of interest from available GWASs in the platform (over 1000 complex traits) or their own GWAS dataset.	No	LD-score regression Automated to run against over 1000 complex traits and diseases simultaneously.	No
Gene-based analysis	MetaXcan, SMR, DEPIC gene-prioritization	MAGMA	No	No
Tissue specificity analysis	DEPICT	MAGMA	No	No
Gene-enrichment analysis	DEPICT	MAGMA	No	No
Mendelian Randomization	GSMR, SMR	No	No	Over 10 different statistical methods to perform Two-sample mendelian randomization (excluding SMR and GSMR).

Usage example

In this section, we present a series of analyses that can be performed in the CTG-VL. As an example, we extracted the GWAS summary statistics for mood swings derived from UK Biobank (available at http://www.nealelab.is/uk-biobank/). This dataset (in addition to over 1000 others) is available for users in the CTG-VL. In this example, we visualize the GWAS results, estimate its genetic correlation with Neuroticism score and Educational Attainment [23], identify enriched genes, pathways and tissues/cell types that are most likely to underlie the associations and determine whether body mass index (BMI) may have a causal relationship with mood swings. A diagram showing the steps to

complete these tasks inside the CTG-VL is displayed in Figure 2. Supplementary Figure 1-7 display screen shots of each step.

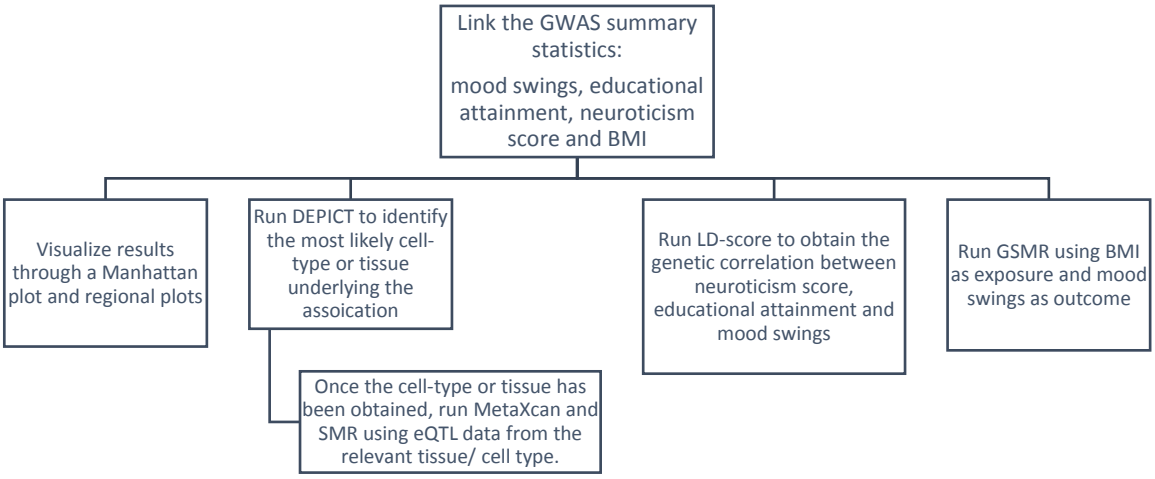

Figure 2. Diagram of the steps to analyse the mood swings GWAS summary statistics.

The first step is to obtain the GWAS summary statistics for mood swings, neuroticism score, educational attainment and BMI. To achieve this, we go into the platform, and after signing in we go to “Data -> Public data”. Here, we look for those traits and click “Add to my data”. Once that is done, these summary statistics can be used throughout the platform. We go to “Visualization -> Manhattan plot”. The CTG-VL implements an interactive Manhattan plot that facilitates the annotation of the SNPs displayed by clicking on them. In addition, the user can highlight all the SNPs that are in LD with the lead-SNP of each locus. Figure 3 shows the Manhattan plot of mood swings with 3 SNPs labelled.

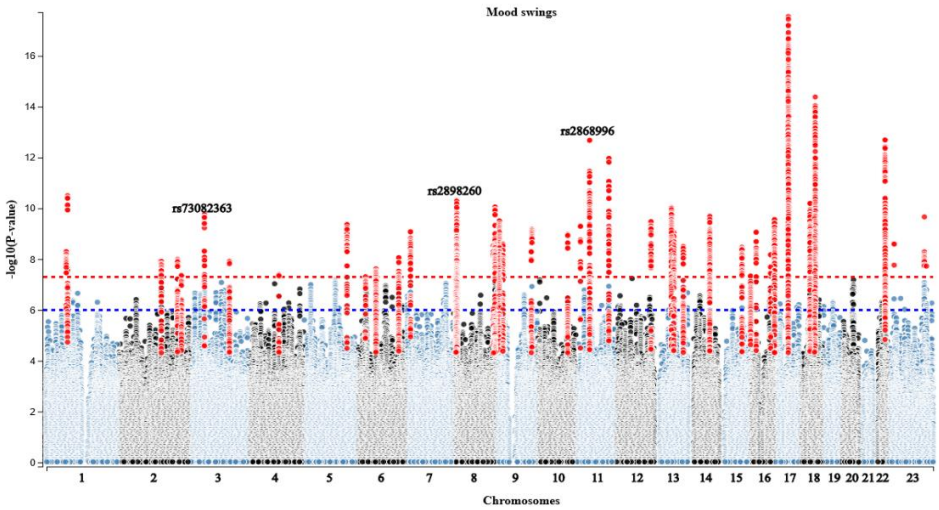

Figure 3. Manhattan plot of mood swings. The CTG-VL enables the user to click on each of the SNPs to obtain the summary statistics, annotate and query the SNP. In addition, the user can highlight SNPs in LD with the lead SNP, as can be seen in red.

In the "Analysis" tab we can then select LD-score, where we estimate the genetic correlation between mood swings and neuroticism score to be rG = 0.87 (s.e. = 0.009). Such high genetic correlation is not surprising since neuroticism score is calculated based on 12 questions, including whether the individual experiences mood swings. Interestingly, we observed a negative genetic correlation between educational attainment and mood swings rG = -0.41 (s.e. = 0.018).

Thereafter, we run DEPICT to obtain the most likely tissues/cell types and biological pathways that may underlie the associations. Tables 2 and 3 summarize the top 5 results of each of these analyses.

Table 2. Top 5 tissues and cell types likely to underlie the associations.

Tissue	Group	P-value	FDR<5%
Hippocampus	Nervous System	0.00000211	Yes
Retina	Sense Organs	0.00000378	Yes
Limbic System	Nervous System	0.00000884	Yes
Cerebral Cortex	Nervous System	0.0000136	Yes
Cerebrum	Nervous System	0.0000173	Yes

Table 3. Top 5 pathways likely to underlie the associations.

Pathway ID	Pathway Description	P-value	FDR<5%
GO:0043523	regulation of neuron apoptotic process	0.0000125	Yes
GO:0051402	neuron apoptotic process	0.0000173	Yes
GO:0070997	neuron death	0.0000209	Yes
GO:0017016	Ras GTPase binding	0.0000370	Yes
GO:0008017	microtubule binding	0.0000565	Yes

We then run MetaXcan and SMR analyses under the tab "Analysis" using gene-expression prediction models and eQTL data derived from the hippocampus (the strongest tissue association derived from DEPICT as shown in Table 2) to prioritize genes that are likely to be associated with mood swings. Top 5 results for MetaXcan and SMR are displayed in Tables 4 and 5 respectively.

Table 4. Top 5 results from MetaXcan using gene-expression prediction models for the hippocampus.

Gene ID	Gene Name	Z-score	P-value	#SNPs used in model
ENSG00000238083	*LRRC37A2*	-7.789	6.73e-15	38
ENSG00000196628	*TCF4*	5.47	4.51e-8	8
ENSG00000213619	*NDUFS3*	5.192	2.08e-7	17
ENSG00000171044	*XKR6*	-5.152	2.57e-7	11
ENSG00000205882	*DEFB134*	5.122	3.03e-7	4

Table 5. Top 5 results from SMR using eQTL data from hippocampus.

Gene ID	Gene Name	Stronger eQTL	Beta	P-value	HEIDI P-value	#SNPs in HEIDI test
ENSG00000214425	*LRRC37A4P*	rs62063676	-0.01	1.10e-10		
ENSG00000214401	*KANSL1-AS1*	rs55974014	0.01	1.35e-10		
ENSG00000238083	*LRRC37A2*	rs2040845	0.01	7.47e-10		
ENSG00000263503	*RP11-707O23.5*	rs55974014	0.01	1.54e-9		
ENSG00000264070	*DND1P1*	rs55974014	0.01	1.94e-9	0.03376701	4

Finally, using GSMR, we assess the causal role of BMI on the risk of mood swings. The results suggest a potential causal effect of BMI on mood swings, with each unit (kg/m^2) higher in BMI leading to a higher risk of mood swings (OR = 1.01, s.e. = 0.006). However, the HEIDI-outlier test result (P-value = 0.053) suggests potential pleiotropic effects of the genetic variants used in the test on both exposure (BMI) and the outcome (mood swings).

Methods

Implementation

The web interface and server are written in Angular 6 and Node.js respectively. Scripts to analyze networks were made with python (2.7) using NetworkX (1.8.1) and Igraph (0.7.1) libraries. Tabix [24] is used to index and query user GWAS data. Graphics were made with in-house d3.js v5 scripts. Genomic annotations and user data are stored at a server in the Queensland Research and Innovation Services Cloud. We use Google's firebase for authentication and meta-data storage. All communications between the CTG-VL platform back- and front-end are encrypted through SSL (Secure Sockets Layer).

Analysis software

The CTG-VL uses PLINK [25] to clump and calculate LD between SNPs. These two functions are mainly used to obtain LD data for the regional plots (LocusTrack), identify SNPs in LD with lead SNPs (those with an association P-value <5e-8) and identify independent loci to be used by DEPICT.

MetaXcan [17] is implemented in the platform with the default options along with gene expression prediction models from PredictDB (http://predictdb.org/) to carry out gene-based analyses. LD-score regression [14, 15] is implemented along with LD-scores derived from 1000 Genomes (https://data.broadinstitute.org/alkesgroup/LDSCORE/) to estimate the heritability and genetic correlation based on GWAS summary statistics. Prior to LD-score analyses, GWAS summary statistics are run through the munge.py utility and merged to HapMap SNPs to ensure that alleles match between different datasets.

DEPICT [16] is implemented to perform gene prioritization, gene-enrichment and identify the tissue/cell type most likely to underlie the associations. DEPICT takes as input independently associated SNPs below a user-specified P-value threshold.

SMR [18] is implemented along eQTL data for 48 cell types derived from GTEx and downloaded from http://cnsgenomics.com/software/smr/#DataResource. GSMR [19] is a function in GCTA [26] and is implemented with the default parameters.

Each of the individual commands to run the analyses are included in the platform.

Data

Currently, genotypes from 1000 Genomes phase 3 [25] are used throughout the platform as a reference for LD. eQTL data from the GTEx Project [10] is used for annotation in regional plots, SMR and MetaXcan. Chromatin states [26] derived from ENCODE [8] and Roadmap Epigenomics Mapping Consortium [9] are used to annotate regional plots. Genes and SNPs position information is based on the GRCh37 human genome assembly. Descriptions of genes are extracted from RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and UniProt (https://www.uniprot.org/).

GWAS summary statistics integrated within the CTG-VL platform are updated regularly, and both the reference and links can be found in (https://genoma.io/data/gwas). GWAS summary statistics are integrated as published by the authors with exception to the summary statistics derived from Neale lab (http://www.nealelab.is/uk-biobank/) where we removed SNPs with a MAF < 0.001 or Minor Allele Count < 25 in Cases (for case-control GWAS) as recommended in their documentation.

Post-GWAS analyses in over 1500 traits

We downloaded 3670 GWAS summary statistics from Neale Lab (http://www.nealelab.is/uk-biobank/) and removed variants as described in the previous sections. We ran LD-score regression against each of these GWAS summary statistics. Those traits with a statistically significant heritability of P-value <0.05 were selected to be analyzed with DEPICT and MetaXcan. As a heuristic, we ran MetaXcan using whole-blood gene-expression prediction models as these models were derived from a larger sample compared to other tissues. To run DEPICT, we clumped the summary statistics with the recommended parameters (LD-window 500Kb, and R^2 < 0.05) and used SNPs with the recommended P-value threshold of <1e-5. For traits with less than 10 independent loci, we relaxed the threshold to P-value <1e-4.

Updates

The CTG-VL is updated on a weekly basis, and this short report may not list/reference all the tools and data in the current version. However, this information can be found on the main page of the platform.

Remarks and Conclusion

We present the first release of the CTG-VL. This platform aims to complement already widely used web applications including LDHub [4], MR-Base [6] and FUMA-GWAS [5] that incorporate some of the functions provided in the CTG-VL. This work was inspired by big efforts from multiple research groups making phenome-wide GWAS summary statistics available (i.e. Neale's lab at the Broad Institute, GeneAtlas [27], and those GWAS summary statistics released along with the SAIGE software [28]) that await to be analyzed further, and more importantly, interpreted.

The CTG-VL intends to be a user-friendly platform that will allow research teams with different expertise to perform common post-GWAS analyses and to aid in the interpretation of this huge amount of data.

The CTG-VL aims to be a community-driven platform where novel analysis tools and data are regularly incorporated as these become available. As such, this report aims to gauge interest from the complex-traits genetics research community.

Acknowledgements

GCP is supported by an Australian Research Council Discovery Early Career Research Award [DE180100976]. ML is supported by a Research Training Program Scholarship from The University of Queensland. PFK is supported by a Research Training Program Scholarship from Queensland University of Technology and QIMR Berghofer Medical Research Institute.

Contributions

GCP designed and developed the platform, implemented workflows, built databases and ran analyses. ML aggregated GWAS summary statistics, implemented workflows and ran analyses. PFK implemented workflows. SDU ran analyses. LFGM implemented workflows. LDH implemented workflows and build databases. Everyone wrote and review this manuscript.

References

1. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.
2. Cloney, R., *Complex traits: Integrating gene variation and expression to understand complex traits.* Nat Rev Genet, 2016. **17**(4): p. 194.
3. 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, 2015. **12**(3): p. e1001779.
4. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2017. **33**(2): p. 272-279.
5. Watanabe, K., et al., *Functional mapping and annotation of genetic associations with FUMA.* Nat Commun, 2017. **8**(1): p. 1826.
6. Hemani, G., et al., *The MR-Base platform supports systematic causal inference across the human phenome.* Elife, 2018. **7**.
7. Cuellar-Partida, G., M.E. Renteria, and S. MacGregor, *LocusTrack: Integrated visualization of GWAS results and genomic annotation.* Source Code Biol Med, 2015. **10**: p. 1.
8. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

9. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
10. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
11. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
12. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Curr Protoc Hum Genet, 2013. **Chapter 7**: p. Unit7 20.
13. Xulvi-Brunet, R. and H. Li, *Co-expression networks: graph properties and topological comparisons.* Bioinformatics, 2010. **26**(2): p. 205-14.
14. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.
15. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.
16. Pers, T.H., et al., *Biological interpretation of genome-wide association studies using predicted gene functions.* Nat Commun, 2015. **6**: p. 5890.
17. Barbeira, A.N., et al., *Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.* Nat Commun, 2018. **9**(1): p. 1825.
18. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.
19. Zhu, Z., et al., *Causal associations between risk factors and common diseases inferred from GWAS summary data.* Nat Commun, 2018. **9**(1): p. 224.
20. Kyoko Watanabe, S.S., Oleksandr Frei, , *A global view of pleiotropy and genetic architecture in complex traits.* bioRxiv, 2018.
21. de Leeuw, C.A., et al., *MAGMA: generalized gene-set analysis of GWAS data.* PLoS Comput Biol, 2015. **11**(4): p. e1004219.
22. Speed, D. and D.J. Balding, *SumHer better estimates the SNP heritability of complex traits from summary statistics.* Nat Genet, 2018.
23. Lee, J.J., et al., *Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals.* Nat Genet, 2018. **50**(8): p. 1112-1121.
24. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files.* Bioinformatics, 2011. **27**(5): p. 718-9.
25. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
26. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.