A comparison of two workflows for regulome and transcriptome-based prioritization of genetic variants associated with myocardial mass

[RUNNING TITLE: Variant prioritization workflow comparison]

Elisabetta Manduchi[1*ǂ], Daiane Hemerich[2*], Jessica van Setten[2], Vinicius Tragante[2], Magdalena Harakalova[2], Jiayi Pei[3], Scott M. Williams[4], Pim van der Harst[5], Folkert W. Asselbergs[2,6,7ǂ], Jason H. Moore[1ǂ]

[1]Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
[2]Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[3]Department of Nephrology and Hypertension, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[4]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Ohio, United States of America
[5]Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.
[6]Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom
[7]Health Data Research UK and Institute of Health Informatics, University College London, London, United Kingdom

[*]These authors contributed equally to this work
[ǂ]Corresponding authors
e-mails: manduchi@pennmedicine.upenn.edu (EM), f.w.asselbergs@umcutrecht.nl (FWA), jhmoore@upenn.edu (JHM); phones: +1 215 573 4408 (EM), +31 88 75 570 6 (FWA), +1 215 573 4411 (JHM)

ORCID:
0000-0002-4110-3714 (EM); 0000-0002-7929-8107 (DH); 0000-0002-4934-7510 (JvS); 0000-0002-8223-8957 (VT); 0000-0002-4835-9544 (SMW); 0000-0002-1692-8669 (FWA); 0000-0002-5015-1099 (JHM)

Abstract

A typical task arising from main effect analyses in a Genome Wide Association Study (GWAS) is to identify Single Nucleotide Polymorphisms (SNPs), in linkage disequilibrium with the observed signals, that are likely causal variants and the affected genes. The affected genes may not be those closest to associating SNPs. Functional genomics data from relevant tissues are believed to be helpful in selecting likely causal SNPs and interpreting implicated biological mechanisms, ultimately facilitating prevention and treatment in the case of a disease trait. These data are typically used post GWAS analyses to fine-map the statistically significant signals identified agnostically by testing all SNPs and applying a multiple testing correction. The number of tested SNPs is typically in the millions, so the multiple testing burden is high. Motivated by this, in this work we investigated an alternative workflow, which consists in utilizing the available functional genomics data as a first step to reduce the number of SNPs tested for association. We analyzed GWAS on electrocardiographic QRS duration using these two workflows. The alternative workflow identified more SNPs, including some residing in loci not discovered with the typical workflow. Moreover, the latter are corroborated by other reports on QRS duration. This indicates the potential value of incorporating functional genomics information at the onset in GWAS analyses.

Keywords: GWAS; functional genomics; SNP preselection; left ventricular mass

Data Availability Statement

No new data were generated in this study. The QRS duration summary statistics for the discovery GWAS are available upon request to the authors of (van der Harst et al., 2016). The replication GWAS data can be obtained from the UK Biobank (www.ukbiobank.ac.uk), upon application. Identifiers for the publicly available ENCODE (www.encodeproject.org), Roadmap Epigenomics (www.roadmapepigenomics.org), and EnhancerAtlas (www.enhanceratlas.org) data used in this work are listed in Supplementary Table 1. For the H3K27ac data on HCM patients, please refer to (Hemerich et al., 2019).

Introduction

Genome Wide Association Studies (GWAS) were introduced over a decade ago as a way to identify genetic risk factors for common human diseases, by analyzing tag Single Nucleotide Polymorphisms (SNPs) that capture the variation at nearby sites in the genome (Hirschhorn and Daly, 2005; Wang et al., 2005). The common method of analysis is to employ univariate statistical tests followed by multiple testing correction, to agnostically identify SNPs that are associated with the trait of interest (Bush and Moore, 2012). The gold standard in the field is replication in independent data sets (NCI-NHGRI Working Group on Replication in Association Studies et al., 2007). While this purely statistical approach is successful at limiting false-positives, it can suffer from inflating the number of false-negatives due to the difficulty in detecting SNPs with small effect sizes under the burden of multiple testing corrections (Williams and Haines, 2011).

In recent years, projects such as ENCODE (ENCODE Project Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics Consortium, 2015) and FANTOM 5 (Andersson et al., 2014; FANTOM Consortium et al., 2014), as well as specific functional genomics efforts by individual laboratories, have made available regulome and transcriptome data sets for a variety of tissues and cell lines, that can be very valuable in the identification of biologically relevant signals from GWAS. The typical approach (which we term Workflow 1 in what follows) employs this functional information post GWAS analyses (Figure 1a). Namely, association tests are first carried out agnostically to identify statistically significant SNPs. Then, functional fine-mapping is applied to identify potential causal variants (Spain and Barrett, 2015). An alternative approach (which we term Workflow 2) consists in utilizing the functional genomics data, obtained from tissues relevant to the trait, to preselect SNPs for association tests (Figure 1b). Since a smaller number of SNPs is tested in Workflow 2, the multiple testing correction is relaxed, potentially enabling the detection of SNPs with smaller effect sizes. We say ‘potentially’ because the less stringent significance threshold in Workflow 2 may not necessarily lead to more results than Workflow 1, since at the same time the set of preselected SNPs to be tested is a smaller candidate set.

In this study, we have compared these two workflows which leverage the functional genomics data either post or prior association analyses for a discovery GWAS data set derived from a large population-based study (van der Harst et al., 2016) of pre-clinical cardiovascular disease (CVD) as measured by left ventricular mass (LVM).

LVM is a good predictor of cardiovascular mortality and morbidity in all genders, races, and ages. Several noninvasive measures of LVM have been developed, and the most commonly applied are electrocardiogram (ECG) measurements of the QRS complex. In van der Harst et al. (2016), GWAS meta-analyses of four such measures led to the identification of 52 associating loci, the majority of which were in non-coding regions of the genome (Hemerich et al., 2019). In this work, we have focused on one of these measurements, namely QRS duration (Verdecchia et al., 1998). For this study, we utilized the summary statistics from the QRS duration GWAS meta-analyses by van der Harst et al. (2016). To assess replication, we utilized the UK Biobank cohort (Sudlow et al., 2015).

We leveraged both public and in-house histone modifications, DNA accessibility, and enhancer prediction data in heart. We included samples from both Hypertrophic CardioMyopathy (HCM) patients (cardiac septum) and controls (Left Ventricle - LV, Human Cardiac Fibroblasts – HCF, and human cardiomyocytes). We used two different approaches to generate 'functional regions' from these data; FILTER 1 where we operationally defined rules for functionality and FILTER 2 where we employed ChromHMM (Ernst and Kellis, 2012), see Methods. We ran our workflow comparison for each of these filters. With either choice, we found a larger number of signals and loci when using the workflow that preselects the SNPs based on the functional regions as compared to performing the statistical testing a priori. Moreover, these loci were consistent with findings in van der Harst et al. (2016), where the latter could leverage full access to that consortium data. At the same time, we imposed consistency with an additional replication data set, not used in van der Harst et al. (2016). Thus, based on these additional constraints, the preselection approach of Workflow 2 displayed better sensitivity without detriment to specificity.

Results

We first analyzed the data with a typical workflow (Workflow 1), namely we first identified statistically significant replicated SNPs, then retrieved their proxy SNPs in high Linkage Disequilibrium (LD; $r^2>0.8$, using http://raggr.usc.edu/ with all European populations), and finally extracted the proxy SNPs harbored in tissue specific functional regions. In order to test different approaches to functional region definition, we did this separately using either FILTER 1 or FILTER 2, described in more details in Methods. We analyzed a total of 2,234,843 SNPs common to the discovery and replication data sets and we used a strict Bonferroni correction (see Methods).

Therefore, the significance threshold in the discovery data set was set to 2.24×10^{-08}, which yielded 666 significant SNPs. We then examined these 666 SNPs in the replication data set (significance threshold=$0.05/666=7.51\times10^{-05}$) and found that 374 replicated, all with the same effect directions as in the discovery data set. We grouped the latter SNPs into 'loci', as in van der Harst et al. (2016). Namely, we sorted the 374 SNPs by increasing p-values and defined the first locus as the 2MB region around the SNP with the lowest p-value (which served as the lead SNP for this locus). If a SNP down this list was within 1Mb from the lead SNP of an already defined locus, it was assigned to that locus, otherwise a new locus was created with that SNP as lead. Under this definition, each locus could have more than one independent signal. With this approach, the 374 SNPs were distributed across 14 loci (Table 1). We then fine-mapped these 374 SNPs by identifying proxies harbored in the functional regions. Tables 2 and 3 indicate (column 'Workflow 1"), for each of the 14 loci, how many of the replicated SNPs remained after fine-mapping employing FILTER 1 and FILTER 2, respectively. After the Workflow 1 fine-mapping with FILTER 1 (respectively FILTER 2), we were left with only eight (respectively seven) of the 14 loci.

In Workflow 2, SNPs are first preselected based on the functional regions, then analyzed in the GWAS. Using FILTER 1, we found 130,836 of the SNPs common to the discovery and replication data sets residing within the deduced functional regions. The conservative Bonferroni corrected threshold in the discovery data set was therefore set to 3.82×10^{-07} (Methods), which yielded 84 significant SNPs. We then examined these 84 SNPs in the replication data set (significance threshold=$0.05/84=5.95\times10^{-4}$) and found that 54 of these SNPs replicated, all with the same effect directions as in the discovery data set. Mapping these 54 SNPs to the 14 loci from Table 1, we found that 10 of those loci harbored relevant SNPs. Moreover, we also found a previously unidentified SNP (rs2840167) defining a new locus on chromosome 2, where by 'new' we mean a locus that was not among the 14 from Table 1. Figure 2 summarizes the comparison between the results from Workflow 1 and Workflow 2, using FILTER 1. Table 2 provides the corresponding details at the locus level. Overall Workflow 2 identified 13 SNPs and three loci that were not identified in Workflow 1. Workflow 1 only identified one SNP (rs2109517) that was not identified in Workflow 2, but this was in a locus that was also identified in Workflow 2. Moreover, the specific p-value of this SNP in the discovery data set was 1.471×10^{-06}, one order of magnitude higher than the p-values of the eight SNPs within the same locus identified by Workflow 2. Summarizing, with FILTER 1, Workflow 1 identified eight loci, as six of the loci from Table 1 did not have any functional SNP in high LD with the significant signals. Workflow 2, on the other hand, identified 11 loci, including all eight discovered with Workflow 1. Indeed, whereas Workflow 1 did not detect any functional SNP in loci chr5:152869040-154869040 and chr7:34404590-36404590, Workflow 2 detected one such SNP in each of these loci. In addition, Workflow 2 identified a new locus. The SNP identified solely by Workflow 2 (rs2840167) and defining the new locus resides in an intron of *CRIM1*, within a region that was operationally identified as a putative active enhancer based on our FILTER 1 criteria. We queried this SNP in HaploReg v4.1 (Ward and Kellis, 2012), where it is also annotated as being within a potential enhancer in heart tissues, including LV. According to PhenoScanner (Staley et al., 2016), this SNP is also associated to forced vital capacity, an interesting related trait. We note that the locus defined by this SNP also contains rs3770900, which was identified as a potential secondary SNP with independent effects on QRS by van der Harst et al. (2016). (Note that both rs2840167 and rs3770900 do not have a small enough p-value in the UK Biobank data set to pass the Workflow 1 replication threshold, but rs2840167 passes the replication threshold in Workflow 2.) All of these elements indicate that the locus defined by rs2840167 is interesting, and we will show below that it is also detected by Workflow2 using the FILTER 2 criteria.

We then examined Workflow 2 using the functional regions derived in FILTER 2. We found that 233,790 of the SNPs common to the discovery and replication data sets resided within these functional regions. The Bonferroni corrected threshold in the discovery data set was therefore set to 2.14×10^{-07}, which yielded 109 significant SNPs. We then examined these 109 SNPs in the replication data set (significance threshold=$0.05/109=4.59\times10^{-04}$) and found that 64 of these SNPs replicated, all with the same effect directions as in the discovery data set. Mapping these 64 SNPs to the 14 loci from Table 1, we recovered 10 of these and we also found six SNPs defining a new locus on chromosome 2 (with lead SNP rs1523787). Supplementary Figure 1 summarizes the comparison between the results from Workflow 1 and Workflow 2, using FILTER 2. Table 3 provides the corresponding details at the locus level. Overall, with the latter filter, Workflow 2 identified 16 SNPs and three loci which were not identified in Workflow 1. Workflow 1 only identified three SNP which were not identified in Workflow 2; rs618472 and rs694808 in locus chr18:41436652-43436652 and rs7526429 in locus chr1:50546140-52546140. Of these two loci, only the former was not also identified in Workflow2. Summarizing, with FILTER 2, Workflow 1 identified seven loci, as the remaining loci from Table 1 did not have any functional SNP in high LD with the significant signals. Only one of these seven loci was not also identified in Workflow 2. Workflow 2, on the other hand, identified 11 loci. Indeed, whereas Workflow 1 did not detect any functional SNP in loci chr1:60897967-62897967, chr5:152869040-154869040, chr7:34404590-36404590 and chr7:115191301-

117191301, Workflow 2 identified functional SNPs in all these loci. Moreover, Workflow 2 identified a new locus. The six SNPs identified solely by Workflow 2 and harbored in the new locus on chromosome 2 were all within 500kb from the lead SNP at this locus, rs1523787. The latter is in high LD (r^2=0.91) with and <10kb from the SNP defining the new locus with FILTER 1 in Workflow 2, indicating that this signal was robust to the different criteria used to build the functional regions.

The two filters that we constructed identified fairly different sets of SNPs harbored within their corresponding functional regions (69,139 in common), and at the SNP level we observed 13 SNPs detected using either filter in Workflow 2, two of which were not detected by Workflow 1 in either case. The latter were rs1003549 in locus chr7:34404590-36404590 and rs2270188 in locus chr7:115191301-117191301. rs1003549 is marked within an enhancer region in heart both by our filters and by HaploReg. rs2270188, in an intron of *CAV2*, a locus corroborated in van der Harst et al. (2016), is marked within a promoter/enhancer region in heart both by our filters and by HaploReg, moreover it alters the binding motif of the cardiac transcription factor p300 (HaploReg). At the locus level, with Workflow 2 both filters detect loci chr3:37767315-39767315, chr6:35622900-37622900, chr6:117667522-119667522, chr1:60897967-62897967, chr5:152869040-154869040, chr12:113793240-1157932401, chr7:34404590-36404590, chr7:115191301-117191301, and the locus on chromosome 2 not discovered with Workflow 1. Thus, nine of the 11 discovered loci using either filter are in common, and these include three loci which were solely discovered by Workflow 2 in either case (chr5:152869040-154869040, chr7:34404590-36404590 and the new locus on chr2).

Discussion

GWAS have successfully identified thousands of *bona fide* associations between SNPs and phenotypes using an agnostic approach. However, various issues have been recognized. The first is that a ‘sentinel’ SNP, i.e. the SNP with the lowest p-value at a given locus, might not necessarily be the actual underlying causal variant, nor is the nearest gene necessarily the functional gene that the SNP is tagging. There is therefore the need to fine-map the GWAS results, i.e. refine the set of potential causal variants, using statistical or data-driven approaches. When functional genomics data are available for tissues relevant to the trait of interest, these can be used to identify SNPs in high LD with the sentinel SNPs and residing in biologically relevant regions. Another issue is that the classical agnostic approach, where univariate tests are run prior to functional fine-mapping, can suffer from inflating the number of false-negatives due to the difficulty in detecting SNPs with small effect sizes under the burden of multiple testing corrections. Motivated by these issues, we have explored the possibility of using tissue specific functional regions to preselect the SNPs to test for association; this approach reduces the number of tested SNPs, thereby relaxing the multiple testing correction penalty. This could potentially enable the discovery of additional SNPs, harbored within functional regions, but there is no guarantee of obtaining additional signals given that the number of candidates is smaller than in the agnostic approach.

We tested how this alternative approach compares to the classical approach in a real data scenario; we leveraged a discovery and a replication GWAS related to QRS duration, a proxy for myocardial mass, relevant to CVD. In this context we observed that the alternative approach yielded several more statistically significant SNPs and overall more replicated loci. Three of the loci were identified solely with the preselection workflow (Workflow 2) regardless of the approach used to define functional regions (FILTER 1 or FILTER 2). Locus chr5:152869040-154869040 is centered around rs13165478, a SNP reported in van der Harst et al. (2016) as associated to QRS duration. Even though this SNP was significant and replicated in the UK Biobank GWAS, it had no proxies residing in a functional region, hence the locus was not detected in Workflow 1. However, Workflow 2 detected this locus via other functional SNPs which were significant and replicated at the thresholds for this workflow. Locus chr7:34404590-36404590 is centered around rs340383, a SNP identified as a potential secondary SNP with independent effects on QRS by van der Harst et al. (2016). Also this SNP had no proxies residing in a functional region, hence the locus was not detected in Workflow 1. However, Workflow 2 detected this locus via other functional SNPs, including rs1003549 that resides within functional regions for either filter. Finally, the last locus discovered only by Workflow 2 on chr 2, as discussed earlier, contains rs3770900, which was identified as a potential secondary SNP with independent effects on QRS by van der Harst et al. (2016). This SNP did not have a low enough p-value in the UK Biobank replication data set. However other functional SNPs in this locus could be detected by the thresholds of Workflow 2.

The fact that the loci we discovered solely with Workflow 2 reflect signals observed in van der Harst et al. (2016), based on full access to consortium data (which we did not have for this work), indicates that these are likely to be relevant to CVD. But we also note that, compared to van der Harst et al. (2016), we added the constraint of replication in the UK Biobank. Moreover, we constrained any discovered SNP to reside within the functional regions defined by our filters. With these additional constraints, these loci could only be identified with the preselection workflow. As with any genomic association study, further validation is necessary to confirm which of the detected signals are truly causal and to elucidate the biological mechanisms and effector genes.

The main focus of this work was not cardiovascular biology, rather to compare two ways of integrating functional genomics data with GWAS summary statistics. This study suggests that incorporating functional genomics data at the onset increases power. However, we recognize that the generalizability of this statement needs to be examined. To this end, future work, by us but also hopefully by others, should include investigating this further for other traits for which both GWAS data (discovery and replication) and functional genomics data for target tissues can be obtained. Public resources such as the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) are now providing summary statistics for several traits. In addition, resources such as the UK Biobank (www.ukbiobank.ac.uk), allow investigators to apply for access to specified genotype and phenotype data from over half a million individuals, enabling direct analyses of GWAS data sets. For functional genomics data, it is crucial to team up with domain experts for the trait of interest so to identify the relevant tissues and generate suitable data sets for the latter, should they not be already available in public resources. Domain expertise is also necessary to evaluate the results of each workflow.

One important issue in using functional genomics data relates to the regions used for filtering. As indicated above, first of all, since these regions are tissue specific, one needs to have a good idea of the involved tissue(s) for a trait. Biological knowledge of the latter can vary considerably depending on the trait. Second, even assuming that the functional genomics data come from an actual related tissue, there are several criteria to define functional regions. In this work, we have used two different methods to derive relevant regions from the functional genomics data; in FILTER 1 we operationally defined functional regions based on several rules, whereas in FILTER 2 we leveraged the ChromHMM software. These filters produced very different sets of SNPs to be examined in Workflow 2. However, at the locus level, the results from the two filters in Workflow 2 were consistent, including three loci only detected by this workflow. We also note that the issue of how to best define functional regions is relevant to both workflows, as ultimately these are the regions used for filtering, whether this is done before or after the association analyses. Given the added value that functional genomics data provide towards elucidating the biological mechanisms at play, a crucial area of future research is the development/refinement of methods that can more precisely identify functional regions based on these data.

Methods

GWAS data sets

For our discovery data set, we obtained the summary statistics (p-values and beta) on QRS duration for ~2.7 million SNPs from van der Harst et al. (2016). As replication data set we used 19,416 12-lead rest ECGs from the UK Biobank (as of April 2018). We extracted QRS information from the ECGs. After excluding individuals using a pacemaker and individuals with heart disease (such as atrial fibrillation, heart failure, myocardial infarction, Wolff-Parkinson-White Syndrome and QRS duration > 120ms), we ran a GWAS analysis on QRS duration using 15,251 individuals that passed quality control filters, with the same methods used on the discovery data, i.e. adjusting for sex, height, BMI, age, chip used (for batch effect) and 10 principal components. We obtained summary statistics for ~ 92.7 million SNPs. We focused our subsequent analyses on 2,234,843 SNPs, obtained by taking the SNPs common to the two data sets, after removing SNPs with inconsistent reported alleles or with ambiguous strand alleles (AT, CG).

Functional genomics data

Supplementary Table 1 lists the public data sets used to define tissue specific functional regions. According to the guidelines presented by the Blueprint Epigenome project (Stunnenberg et al., 2016; found at http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch38), we used broad ChIP-seq peaks for histone modifications H3K27me3, H3K36me3, H3K4me1, H3K9me3 and narrow peaks for H3K27ac and H3K4me3.

We also included H3K27ac Chip-seq generated on human myocardial samples from fourteen HCM patients as described in (Hemerich et al., 2019).

Using these data, we inferred relevant functional regions in two different ways (FILTER 1 and FILTER 2), as described below, using hg19 coordinates. BED file manipulations were performed with bedtools v2.27.1 (Quinlan and Hall, 2010) and bedops v2.4.35 25 (Neph et al., 2012).

Definition of functional regions: FILTER 1

Similarly to (Manduchi et al., 2018), we operationally defined active enhancers, promoters and exons in the relevant heart cellular context as described below, and used the union of these three types of regions as filter. When multiple peak files were available for the same mark, they were first merged.

Putative active enhancers

1. Take overlaps between the H3K4me1 and H3K27ac peaks.
2. Flank DNase peaks by 150bp on each side.
3. Take regions from (2) which have at least 1bp overlap with regions from (1).
4. Merge regions from (3) with enhancers for LV downloaded from EnhancerAtlas (Gao et al., 2016).

Putative active promoters

1. Flank DNAse peaks by 150bp on each side.
2. Take regions from (1) which have at least 1bp overlap with H3K4me3 peaks.
3. Merge regions from (2) with 5000 bp regions centered at the transcription start sites (TSS) of genes expressed in LV according to GTEx v7 (GTEx Consortium et al., 2017).
4. Take all regions from 1000bp upstream to 500bp downstream of GeneCode v19 TSS (Frankish et al., 2019), downloaded from the UCSC Genome Browser (Hinrichs et al., 2006).
5. Take all regions in (4) which overlap with regions from (3).

Putative active exons

Take all exons from the transcripts corresponding to the putative active promoters defined above.

Definition of functional regions: FILTER 2

We used the DNase and the histone modification data (for six marks) with ChromHMM v1.17 to model 10 chromatin states. We used as functional regions those corresponding to six of these states which, based on signature, appeared to be potential active enhancer, promoters, exons or silenced regions. Prior to running LearnModel from

ChromHMM, we separately binarized the data for narrow peaks and broad peaks. According to manual recommendations, for narrow peaks, we ran BinarizeBed (from ChromHMM) with the –peak option and with input the peak files from Supplementary Table 1. For broad peaks, we ran BinarizeBed without the –peak option and with input the alignment files in bed format (listed by GEO accession in Supplementary Table 1).

Identification of statistically significant replicated SNPs

We used a significance threshold of 0.05. For a given input set of M candidate SNPs, we first identified those with p-value less than $0.05/M$ in the discovery dataset, hence applying a strict Bonferroni multiple testing correction (this is conservative if the M SNPs are not independent). Then we tested only the resulting N significant SNPs in the replication data set, identifying those with p-value less than $0.05/N$ in the latter (again applying a Bonferroni correction, but the number of tests is smaller). Any SNP passing this test with the same direction of effect in the two data sets was considered significantly replicated. In Workflow 1, the input for the discovery phase consisted of all 2,234,843 SNPs common to the two GWAS described above. In Workflow 2, the input for the discovery phase consisted of the common SNPs which were located in functional regions defined by FILTER 1 or FILTER 2. Figure 1 illustrates the procedure for both workflows.

Acknowledgements

Funding for this work was provided by the National Institutes of Health (NIH) grant LM010098. This research has been conducted using the UK Biobank Resource under Application Number 24711. FWA is supported by UCL Hospitals NIHR Biomedical Research Centre. This work was also partly supported by the Netherlands Cardiovascular Research Initiative - an initiative with support of the Dutch Heart Foundation, CVON: The Netherlands CardioVascular Research Committee, CVON2014-40 DOSIS (The Netherlands).

References

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature, 507*, 455-461. *doi:* 10.1038/nature12787

Bush WS, Moore JH (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol., 8*, e1002822. *doi:* 10.1371/journal.pcbi.1002822

Ernst J., Kellis M. (2012). ChromHMM: automating chromatin state discovery and characterization. *Nat Methods, 9*, 215-216. *doi:* 10.1038/nmeth.1906

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489,* 57-74. *doi:* 10.1038/nature11247

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, et al. (2014). A promoter-level mammalian expression atlas. *Nature, 507,* 462-70. *doi:* 10.1038/nature13182

Frankish A, Diekhans M2, Ferreira AM2, Johnson R, Jungreis I, Loveland J, et al. (2019). *Nucleic Acids Res., 47,* D766-D773. *doi:* 10.1093/nar/gky955

Gao T, He B, Liu S, Zhu H, Tan K, Qian J (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics, 32,* 3543-3551. *doi:* 10.1093/bioinformatics/btw495

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature, 500,* 204-213. *doi:* 0.1038/nature24277

Hemerich D, Pei J, Harakalova M, van Setten J, Boymans S, Boukens BJ, et al. (2019). Integrative Functional Annotation of 52 Genetic Loci Influencing Myocardial Mass Identifies Candidate Regulatory Variants and Target Genes. *Circ Genom Precis Med., 12,* e002328. *doi:* 10.1161/CIRCGEN.118.002328

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. (2006). The UCSC Genome Browser Database: update. *Nucleic Acids Res., 34,* D590-8. *doi:* 10.1093/nar/gkj144

Hirschhorn JN, Daly MJ (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet., 6*, 95-108. *doi:* 10.1038/nrg1521

Koch L (2014). Disease genetics: insights into missing heritability. *Nat Rev Genet., 15,* 218. *doi:* 10.1038/nrg3713

Manduchi E, Williams SM, Chesi A, Johnson ME, Wells AD, Grant SFA, Moore JH (2018). Leveraging epigenomics and contactomics data to investigate SNP pairs in GWAS. *Hum Genet., 137,* 413-425. *doi:* 10.1007/s00439-018-1893-0

Moore JH, Asselbergs FW, Williams SW (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics, 26,* 445-455. *doi:* 10.1093/bioinformatics/btp713

NCI-NHGRI Working Group on Replication in Association Studies et al. (2007). Replicating genotype-phenotype associations. *Nature, 447,* 655-660. *doi:* 10.1038/447655a

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics, 28*, 1919-1920. *doi:* 10.1093/bioinformatics/bts277

Quinlan AR, Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26,* 841-842. *doi:* 10.1093/bioinformatics/btq033

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518,* 317-330. *doi:* 10.1038/nature14248

Spain LS, Barrett JC (2015). Strategies for fine-mapping complex traits. *Hum Mol Genet., 24,* R111–R119. *doi:* 10.1093/hmg/ddv260

Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics, 32,* 3207-3209. *doi:* 10.1093/bioinformatics/btw373

Stunnenberg HG, International Human Epigenome Consortium, Hirst M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell, 167,* 1145-1149. *doi:* 10.1016/j.cell.2016.11.007

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al (2015). 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. *doi:* 10.1371/journal.pmed.1001779

van der Harst, P, van Setten J, Verweij N, Vogler 4, Franke L, Maurano MT, et al. (2016). 52 Genetic Loci Influencing Myocardial Mass. *J Am Coll Cardiol,, 68,* 1435-1448. *doi:* 10.1016/j.jacc.2016.07.729

Verdecchia P, Schillaci G, Borgioni C, Ciucci A, Gattobigio R, Zampi I, Porcellati C (1998). Prognostic value of a new electrocardiographic method for diagnosis of left ventricular hypertrophy in essential hypertension. *J Am Coll Cardiol,, 31,* 383-390. *pmid:* 9462583

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet., 6,* 109-118. *doi:* 10.1038/nrg1522

Ward LD, Kellis M (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res., 40,* D930-934. *doi:* 10.1093/nar/gkr917

Williams SM, Haines JM (2011). Correcting away the hidden heritability. *Ann Hum Genet., 75,* 348-350. *doi:* 10.1111/j.1469-1809.2011.00640.x

[dataset] van der Harst, P, van Setten J, Verweij N, Vogler 4, Franke L, Maurano MT, et al. (2016). 52 Genetic Loci Influencing Myocardial Mass. *J Am Coll Cardiol,, 68,* 1435-1448. *doi:* 10.1016/j.jacc.2016.07.729

[dataset] Hemerich D, Pei J, Harakalova M, van Setten J, Boymans S, Boukens BJ, et al. (2019). Integrative Functional Annotation of 52 Genetic Loci Influencing Myocardial Mass Identifies Candidate Regulatory Variants and Target Genes. *Circ Genom Precis Med., 12,* e002328. *doi:* 10.1161/CIRCGEN.118.002328

Table 1

Loci corresponding to the 374 significant replicated SNPs from Workflow 1 (prior to any functional fine-mapping). Coordinates refer to hg19. The lead SNP for each locus is indicated, together with its unadjusted p-values in the discovery and replication data sets. Loci are listed by increasing p-value of their lead SNP in the discovery data set.

Locus	**Lead SNP**	**Discovery p**	**Replication p**
chr3:37767315-39767315	rs6801957	6.90E-40	3.27E-15
chr6:35622900-37622900	rs1321311	1.03E-37	7.19E-15
chr6:117667522-119667522	rs11153730	7.44E-29	1.71E-08
chr1:60897967-62897967	rs2207790	6.71E-19	6.81E-06
chr5:152869040-154869040	rs13165478	8.06E-19	6.95E-16
chr12:113793240-115793240	rs883079	4.58E-16	3.88E-07
chr10:113505465-115505465	rs7918405	1.05E-14	3.81E-08
chr18:41436652-43436652	rs10853525	1.41E-14	1.53E-09
chr1:50546140-52546140	rs17391905	1.07E-11	4.65E-05
chr17:63312463-65312463	rs12940610	1.08E-11	6.02E-05
chr13:73513122-75513122	rs728926	5.60E-11	1.26E-05
chr7:34404590-36404590	rs340383	3.72E-10	2.13E-05
chr7:115191301-117191301	rs11773845	7.50E-10	1.66E-06
chr7:45640900-47640900	rs6968945	5.14E-09	5.04E-05

Table 2

Comparison of results from Workflow 1 and Workflow 2 for the functional regions defined by FILTER 1. For each of the 14 loci from Table 1 and the new locus (gray cells), the number of SNPs identified by each workflow and those identified by both are indicated. The SNPs identified solely by each one of the two workflows are also listed.

Locus	Workflow 1	Workflow 2	Common	Worfkflow 1 only	Workflow 2 only
chr3:37767315-39767315	20	20	20	none	none
chr6:35622900-37622900	3	3	3	none	none
chr6:117667522-119667522	13	15	13	none	rs9489449
					rs3734382
chr1:60897967-62897967	1	1	1	none	none
chr5:152869040-154869040	0	1	0	none	rs17116165
chr12:113793240-115793240	2	2	2	none	none
chr10:113505465-115505465	0	0	0	none	none
chr18:41436652-43436652	1	1	1	none	none
chr1:50546140-52546140	0	0	0	none	none
chr17:63312463-65312463	0	0	0	none	none
chr13:73513122-75513122	0	0	0	none	none
chr7:34404590-36404590	0	1	0	none	rs1003549
chr7:115191301-117191301	1	8	0	rs2109517	rs2191502
					rs8713
					rs6466587
					rs2270188
					rs6466579
					rs6867
					rs1049314
					rs9920
chr7:45640900-47640900	1	1	1	none	none
chr2:35683316-37683316	NA	1	0	NA	rs2840167

Table 3

Comparison of results from Workflow 1 and Workflow 2 for functional regions defined by FILTER 2. For each of the 14 loci from Table 1 and the new locus (gray cells), the number of SNPs identified by each workflow and those identified by both are indicated. The SNPs identified solely by each one of the two workflows are also listed.

Locus	Workflow 1	Workflow 2	Common	Worfkflow 1 only	Workflow 2 only
chr3:37767315-39767315	22	24	22	none	rs12491987 rs6599210
chr6:35622900-37622900	3	4	3	none	rs12207548
chr6:117667522-119667522	8	8	8	none	none
chr1:60897967-62897967	0	2	0	none	rs9436640 rs2103883
chr5:152869040-154869040	0	2	0	none	rs11167682 rs7706345
chr12:113793240-115793240	9	9	9	none	none
chr10:113505465-115505465	1	1	1	none	none
chr18:41436652-43436652	2	0	0	rs618472 rs694808	none
chr1:50546140-52546140	6	5	5	rs7526429	none
chr17:63312463-65312463	0	0	0	none	none
chr13:73513122-75513122	0	0	0	none	none
chr7:34404590-36404590	0	2	0	none	rs1003549 rs2075048
chr7:115191301-117191301	0	1	0	none	rs2270188
chr7:45640900-47640900	0	0	0	none	none
chr2:35673773-37673774	NA	6	0	NA	rs1523787 rs7562790 rs888083 rs12476515 rs2252032 rs3770781

Figure Legends

Figure 1. Abstract outline of the two workflows compared in this study. (a) In Workflow 1, the typical approach, statistically significant replicated SNPs are identified first in the GWAS analyses, then fine-mapped based on tissue specific functional regions. (b) In Workflow 2, tissue specific functional regions are used to preselect the SNPs to be included in the GWAS analyses.

Figure 2. Summary results for Workflow 1 (a) and Workflow 2 (b), for the discovery and replication GWAS on QRS duration examined in this work, using FILTER 1 to define functional regions.