ProGeM: A framework for the prioritisation of candidate causal genes at molecular quantitative trait loci

David Stacey[1,*], Eric B. Fauman[2], Daniel Ziemek[3], Benjamin B. Sun[1], Eric L. Harshfield[1], Angela M. Wood[1], Adam S. Butterworth[1], Karsten Suhre[4], and Dirk S. Paul[1,*]

[1] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[2] Pfizer Worldwide Research & Development, Genome Sciences & Technologies, Cambridge, MA, USA
[3] Pfizer Worldwide Research & Development, Inflammation & Immunology, Berlin, Germany
[4] Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Doha, Qatar

* To whom correspondence should be addressed. Tel: +44 (0)1223 747217; Email: ds763@medschl.cam.ac.uk. Correspondence may also be addressed to. Tel: +44 (0)1223 761918; Email: dsp35@medschl.cam.ac.uk.

ABSTRACT

Quantitative trait locus (QTL) mapping of molecular phenotypes such as metabolites, lipids and proteins through genome-wide association studies (GWAS) represents a powerful means of highlighting molecular mechanisms relevant to human diseases. However, a major challenge of this approach is to identify the causal gene(s) at the observed QTLs. Here we present an analysis framework for the "**Pr**ioritisation **o**f candidate causal **Ge**nes at **M**olecular QTLs" (ProGeM), which incorporates biological domain-specific annotation data alongside genome annotation data from multiple repositories. We assessed the performance of ProGeM using a reference set of 213 previously reported and extensively curated metabolite QTLs. For 98% of these loci (*n*=209), the expert-curated gene was one of the candidate causal genes prioritised by ProGeM. Systematic benchmarking analyses revealed that 70% (*n*=150) of the candidate causal genes were nearest to the sentinel variant at the investigated molecular QTLs, indicating that genomic proximity is the most reliable indicator of true positive causal genes. In contrast, *cis*-gene expression QTL data led to three false positive candidate causal gene assignments for every one true positive assignment. Finally, we provide evidence that these conclusions may also apply to *cis*-protein QTLs and loci associated with complex diseases. ProGeM is freely available via GitHub.

INTRODUCTION

With the continued application of genome-wide association studies (GWAS) to complex human disease aetiology (1-4), the rapid rate of susceptibility locus discovery is far outstripping the rate at which we are able to elucidate the biological mechanisms underlying the identified loci. This represents a major bottleneck to translational progress. Quantitative trait locus (QTL) mapping of molecular phenotypes (also referred to as endophenotypes) provides a powerful means to functionally annotate and characterise GWAS signals for complex traits in a high-throughput manner. This approach has been pioneered with the use of transcriptomic data to identify gene expression QTLs (eQTLs) (5-9). Recent technological advances have also made it feasible to identify thousands of methylomic (meQTL) (10,11), proteomic (pQTL) (12,13) (BioRxiv: doi: https://doi.org/10.1101/134551), lipidomic (14), and metabolomic (mQTL) (15-17) QTLs. This catalogue of molecular QTLs, cutting

across multiple “omic” modalities, can be readily queried to elucidate the functional impact of disease-associated variants not only on the abundance of transcripts, but also proteins, lipids, and metabolites.

A key challenge with these data relates to the issue of linking the observed molecular QTLs to specific causal genes. Accurate molecular QTL-gene assignments are critical for the meaningful interpretation of the biology underlying GWAS signals and the subsequent design of appropriate experimental follow-up work. The traditional approach has been to assign QTLs to the gene that is located nearest to the sentinel variant. However, this “nearest gene” approach does not take into account the complex linkage disequilibrium (LD) structure of the genome, which usually spans multiple genes, as well as the potential of regulatory elements that can exert long-range effects on distal genes (i.e., >1 megabase (mb)) (16,18). A notable example that illustrates the limitations of the approach is provided by the *FTO* gene locus, associated with body mass index. The causal variants at this locus locate in a genic enhancer in *FTO* that was shown to have long-range control (~1.2mb) of the homeobox regulatory genes *NRX3* and *NRX5* (19).

There are now several web-tools to facilitate the identification of genes most likely to be impacted functionally by either the sentinel or proxy variants tagging a molecular QTL. For example, tools such as the Single Nucleotide Polymorphisms Annotator (SNiPA) and the Functional Mapping and Annotation of GWAS tool (FUMA) integrate various positional, regulatory, and *cis*-eQTL datasets, enabling the identification of candidate causal genes in a more informed fashion (20,21). Nevertheless, a major challenge relating to these automated tools lies in being able to return a high proportion of “true positive” causal genes (i.e., high sensitivity) and a low number of accompanying “false positive” genes (i.e., high specificity). Given the small number of trait- or disease-associated variants that have been unequivocally assigned to known causal genes, the sensitivity of these tools has not yet been assessed. However, these tools are typically conservative in that they highlight several candidate causal genes at most QTL, and so it is clear that they are hampered by low specificity.

To this end, increased specificity can be achieved by performing a comprehensive literature review to determine which candidate genes have previously been linked to the molecular phenotype in question, though this approach is very time-consuming as molecular QTL studies can yield hundreds of independent loci. An alternative approach involves applying preselected criteria that likely lead to an increase in true positive causal genes. For example, candidate genes that contain a protein-altering (i.e., missense, frame-shift, nonsense) sentinel or proxy variant can be prioritised above others on the basis that these variants are more likely to have an effect on gene or protein function (22). However, the selection of such criteria needs to be determined in a rigorous and objective manner. Irrespective of the analytical approach, candidate genes need to be validated in experimental studies to ascertain causality.

Here we present an intuitive framework and accompanying R script (https://github.com/ds763/ProGeM) for the Prioritisation of candidate causal Genes at Molecular QTLs (ProGeM). Consistent with existing tools, ProGeM leverages positional and *cis*-eQTL data to prioritise genes most likely to be impacted functionally by variants tagging the molecular QTL. In addition, ProGeM integrates information from biological domain-specific annotation data from multiple databases to prioritise genes involved in the biological mechanisms that regulate the molecular phenotype in question. In this way, ProGeM is able to harness both literature- and experimental-derived information in a quick and efficient manner. Crucially, and in contrast to existing tools, we have also quantified the sensitivity and specificity of ProGeM using two molecular QTL datasets (i.e.,

213 mQTLs, 561 *cis*-pQTLs) for which each QTL has been assigned an expert-curated, high-confidence causal gene as a “gold-standard” comparison. Informed by these unique datasets, we make recommendations as to which annotation criteria may be most informative when aiming to identify candidate causal genes at molecular QTLs.

MATERIAL AND METHODS

Generation of candidate causal gene reference datasets

mQTL dataset. Biochemistry has a rich history, predating genome-wide association studies by a century. This makes metabolite GWAS special in that the functional studies of genes at a locus, which would normally follow a genetic study, have frequently already been done.

Between 2007 and 2016, there were 109 papers reporting results of one or more metabolite QTL GWAS. Suitable traits were identified largely through a manual review of all entries from the GWAS catalogue tagged with the EFO term “measurement” (EFO_0001444) or any descendants of the term. This analysis focused on small molecules, ions, metabolites, vitamins and other biomolecules not directly encoded by genes such as mRNA or proteins. The source tissue was most often plasma or serum although studies of urine and cerebrospinal fluid (CSF) have also been reported. Where available, full summary statistics for the identified studies were downloaded and peak-pruned to identify sentinel SNPs at least 1mb apart. This was done for datasets generated by the Global Lipids Genetics Consortium (i.e. LDL cholesterol, HDL cholesterol, total cholesterol and total triglycerides), Meta-Analyses of Glucose and Insulin-related traits Consortium (i.e. fasting glucose) and TwinsUK/KORA (16), with a total of 78 metabolites with $p \leq 5\times10^{-8}$. Before clustering, there were 2,808 sentinel SNPs ($p \leq 5\times10^{-8}$) covering 250 distinct metabolites from these 109 studies. These variants were clustered into 497 loci by collapsing variants closer than 500kb unless there was a compelling biochemical reason to separate the associated metabolites. For example, nine sentinel SNPs for branched chain amino acids and related metabolites near *PPM1K* were clustered together, and 21 sentinel SNPs for urate, uric acid and urea near *ABCG2* were clustered. However, these two groups were not clustered further, even though they are only 150kb apart because the metabolites are not tightly linked biochemically and each cluster has its own very credible causal gene. Within each cluster, the variant with the strongest *p*-value (across all related metabolites) was retained.

For each of the 497 locus-metabolite pairs, all protein-coding genes within 1mb of the sentinel variant were considered. This generated a median of 20 genes per locus (range 4 to 92). While the final selection of the likely causal gene was performed manually following an expert review of the literature, a variety of approaches was employed to guide the literature review. Text-mining was used to identify papers mentioning the gene, gene product or synonyms of each and the metabolite or synonyms of the metabolite. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway databases were mined for networks linking each gene at a locus to the associated metabolite. In addition, the original GWAS papers often speculate on causal genes for each metabolite and these were considered as well. Many metabolites are so distinct that only a handful of genes in the whole genome have ever been discussed in relation to them. Examples include 5-oxoproline, here linked to the *OPLAH* gene that encodes 5-oxoprolinase (23), and manganese, here linked to the *SLC30A10* gene that encodes a manganese transporter (24).

For each of the 497 gene-metabolite pairs, we attempted to identify the earliest publication conclusively linking the gene product to the exact metabolite reported or a biochemically-similar metabolite. Preference was given to evidence for the human gene, to non-genetic data and to experimental work done before the publication of the GWAS. The publications reporting the experimental validation for the causal genes are presented in **Supplementary Table S1**, listed under "Evidence Source (PMID)".

cis-pQTL dataset. This dataset was derived from a recent large-scale pQTL study (BioRxiv: doi: https://doi.org/10.1101/134551). Briefly, the dataset consisted of 3,301 healthy individuals of European descent, who had been randomly selected from a pool of ~50,000 participants of the INTERVAL study (25). Plasma protein levels were measured using the SOMAscan platform (SomaLogic, Inc., Boulder, Colorado, US) comprising 4,034 distinct aptamers (SOMAmers) covering 3,623 proteins (or protein complexes). Genotyping was performed using the Affymetrix Axiom UK Biobank genotyping array (Santa Clara, California, US), assaying ~830,000 variants. Variants were imputed using a combined 1000 Genomes Phase 3-UK10K reference panel, which yielded a total of ~10.5 million variants for pQTL analyses after stringent QC filtering. Overall, the authors found a total of 1,928 significant ($p<1.5\times10^{-11}$) genetic variant-protein associations involving 1,479 proteins and 765 unique genomic loci (BioRxiv: doi: https://doi.org/10.1101/134551). Of these 1,928 associations, 554 were *cis*-associations (i.e., sentinel variant within 1mb of the gene encoding the corresponding protein) and the remaining 1,374 were *trans*-associations.

In order to benchmark our candidate causal gene identification strategy, we utilised only the *cis*-pQTL data, for which we hypothesised that the causal gene at a given *cis*-pQTL ought to be the gene that encodes the associated protein. To convert the 554 *cis*-pQTLs into a high-confidence set of sentinel variant-causal gene assignments, we first decomposed the *cis*-pQTLs into 588 sentinel variant-SOMAmer *cis*-associations. We then removed nine sentinel variants with associations originating from SOMAmers known to target more than a single protein due to paralogous sequences. We also removed 16 associations with SOMAmers that led to duplicate (or more) protein associations for the same sentinel variant. Finally, we removed two additional associations for which the same sentinel variants were associated with distinct protein isoforms encoded by single genes. Thus, we used a set of 561 high-confidence sentinel variant-causal gene assignments (**Supplementary Table S2**) for the purposes of benchmarking the bottom-up component of ProGeM.

Coronary artery disease dataset. This dataset was derived from recently published GWAS of coronary heart disease (CAD) (3,26,27). At 78 independent GWAS loci, the authors assigned each association locus to one or more plausible biological candidate causal genes and to a causal biological pathway (where possible). These candidate genes have largely been defined by epidemiologically-identified risk factors, e.g., *IL6R* (IL6 inflammatory signalling); *LPA*, *LIPA* and *LDLR* (LDL cholesterol and lipoprotein(a)); etc. For the purposes of our study, we first removed those loci that had been assigned to more than one causal gene, and we also removed two additional loci that were assigned to non-protein coding genes. This left us with a list of 56 independent CAD loci, each assigned to a single protein-coding gene (**Supplementary Table S3**).

Proxy variant selection

We selected proxies for each sentinel variant based on an LD threshold of $r^2\geq0.8$. For the mQTL dataset, proxies were extracted from the 1000 Genomes Project (EUR Super Population) data using PhenoScanner, which is a curated database of publicly available results from large-scale

genetic association studies (28). For the *cis*-pQTL dataset, proxies were derived directly from the genotype data of the participants (BioRxiv: doi: https://doi.org/10.1101/134551) (see above).

Annotation of sentinel and proxy variants

All sentinel and proxy variants were annotated using the Ensembl Variant Effect Predictor (VEP) (v83) on GENCODE transcripts (v19) for GRCh37 (29). Annotations were generated using the "per gene" option, which considers variant annotations across all genes and transcripts and selects the most severe consequence per gene with an arbitrary selection of the corresponding transcript. In particular, we made use of the IMPACT rating provided by VEP, which assigns input variants to one of four overarching functional categories as follows: (i) high impact: "the variant is assumed to have high (disruptive) impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay" (i.e., frameshift variant, start-lost variant); (ii) moderate impact: "a non-disruptive variant that might change protein effectiveness" (i.e., missense variant, inframe deletion); (iii) low impact: "assumed to be mostly harmless or unlikely to change protein behaviour" (i.e., synonymous variant, 3'-untranslated region variant); and (iv) modifier impact: "usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact" (i.e., intergenic variant, intronic variant).

Identification of candidate causal genes

Bottom-up component. We used the GenomicRanges suite of R packages (30) to extract (i) the three nearest protein-coding genes to each sentinel variant, and (ii) any LD range-overlapping genes from a GRCh37 gene model based on a GTF file ("Homo_sapiens.GRCh37.82.gtf") retrieved from Ensembl (31). LD ranges for each sentinel variant were defined as the range between the genomic coordinates (GRCh37) of the left- and right-most proxy variants (+/-5kb). In cases where the sentinel had no proxies, the coordinates of the sentinel variant (+/-5kb) were taken as the LD range. We also extracted significant *cis*-eQTL target genes of either sentinel or proxy variants from the *cis*-eQTL data prepared by the Genotype-Tissue Expression (GTEx) project (5) (v6p), across all tissues assayed (n=44). Significant *cis*-eQTLs were defined by beta distribution-adjusted empirical p-values using a false discovery rate (FDR) threshold of 0.05 (see http://www.gtexportal.org/home/documentationPage for details).

Top-down component. mQTL sentinel variant-flanking genes (i.e., transcription start site (TSS) within +/-500kb of a sentinel) were identified using GenomicRanges (30) and the same Ensembl GTF file as above. Top-down candidates were then identified by cross-referencing the resultant list of sentinel-flanking genes against a list of known metabolic-related genes derived from five open-source databases (**Supplementary Table S4**).

Statistical analyses

Calculating specificity. We calculated the overall specificity of ProGeM as:

$$TN \div (TN + FP)$$ [TN: true negatives, FP: false positives]

whereby maximal specificity is indicated by specificity = 1, and where TN was comprised of:

$$\Sigma LG - (TP + FP) - FN$$ [LG: local protein-coding genes residing +/-500kb

from a sentinel variant, TP: true positives, FN: false negatives]

To compare the specificity of the bottom-up and top-down components, as well as the concurrent candidate gene sets, TN was comprised of:

$$\Sigma PG - (TP + FP) - FN$$ [PG: ProGeM candidates for each sentinel variant]

Enrichment analyses. Enrichment analyses were performed using Fisher's exact tests, with the relevant background gene sets consisting of all remaining candidate causal genes (both bottom-up and top-down) across either the mQTL or the *cis*-pQTL dataset as appropriate.

Software for analyses

All analyses described in this study were performed using R v3.3.2 and Bioconductor v3.3.

RESULTS

Conceptual framework of ProGeM

We have devised a general framework for the **Pr**ioritisation **o**f candidate causal **Ge**nes at **M**olecular QTLs (ProGeM) (**Figure 1**). This framework is based on the assumption that in order for a gene to be a causal candidate at a molecular QTL, it must fulfil at least one of the following two requirements: (i) the gene must exhibit altered structure, abundance or function as a result of the sentinel or proxy variants at the QTL, or (ii) the gene must be directly involved in the molecular mechanisms and/or biological processes that influence the trait in question. Accordingly, ProGeM is comprised of a "bottom-up" and a "top-down" component that prioritises candidate causal genes from the perspective of the genetic variant and the phenotype, respectively.

Bottom-up component. For the bottom-up component, we utilise three complementary methods to identify plausible candidate causal genes based on (i) their proximity to the LD range (**Material and Methods**), (ii) their proximity to the sentinel variant, and (iii) whether their mRNA expression levels are impacted by either the GWAS sentinel or any corresponding proxy variants (**Figure 1**). The former two methods are designed to capture candidate causal genes that are proximal to the association signal, whereas the latter enables inclusion of more distal candidate genes. Any gene that meets at least one of these criteria is included in the list of bottom-up candidate genes. In addition, those candidate causal genes that contain either a sentinel or proxy variant of high or moderate impact on gene function (**Material and Methods**) are annotated as such.

Top-down component. For the top-down component, we first identify all genes that reside within a pre-defined genomic window either side of the sentinel variant. Various open-source databases are then referenced to determine whether any of these genes have previously been implicated in the regulation of the molecular phenotype in question, thereby constituting the top-down candidates (**Figure 1**). The type of databases referenced, and the way in which they are queried (**Supplementary Table S4**), depends on the nature of the molecular phenotype in question (e.g., the abundance of proteins, metabolites or lipids, etc.). For the purposes of this study, we extracted a list of metabolic-related genes from five databases: (i) Gene Ontology (GO), (ii) Kyoto Encyclopedia of Genes and Genomes (KEGG), (iii) Mouse Genome Informatics (MGI), (iv) Orphanet, and (v)

Reactome (**Figure 1**; **Supplementary Table S4**). We have made this list available at GitHub (https://github.com/ds763/ProGeM). Lastly, the top-down candidate genes are assigned a score ranging between 1 and 5 to reflect the number of times they are reported in the databases.

Framework integration. Finally, the lists of bottom-up and top-down candidate genes for each identified QTL are integrated by ProGeM to determine whether any genes are identified by both independent approaches. Any concurrent candidate genes are then designated the most likely causal genes given that they fulfil both of the aforementioned requirements for a candidate causal gene.

ProGeM implementation and parameter selection

ProGeM is implemented within the R statistical environment as a configurable .R script, which is freely available at GitHub along with a .readme file describing the necessary input and resultant output files (https://github.com/ds763/ProGeM).

The parameters used by ProGeM can be adjusted based on the type of molecular QTL data and the research question provided by the user. Specifically, these parameters include (i) the number of nearest genes to each sentinel variant that ProGeM should consider to be candidate causal genes ("number of nearest genes"; default=3); (ii) the size of the genomic window around each sentinel variant from which candidate genes are reported ("distance"; default=500 kilobases (kb)); (iii) the threshold ProGeM should use to select proxies from a user-supplied file ("r^2 threshold"; default≥0.8); and (iv) the threshold ProGeM should use to select *cis*-eQTL target genes as candidate causal genes ("*cis*-eQTL *p*-value threshold"=default: beta distribution-adjusted empirical *p*-values using a FDR threshold of 0.05, see http://www.gtexportal.org/home/documentationPage for details).

In order to determine how the ProGeM output is impacted by changing various parameters, we applied ProGeM to a list of 213 literature-derived mQTLs where each QTL has been assigned a high-confidence causal gene (**Material and Methods; Supplementary Table S1**). We ran an additional iteration of ProGeM after each parameter change (while leaving all others in the default setting), and then measured both sensitivity and specificity. These metrics are defined by comparison of the genes identified by ProGeM to the putative causal gene defined by expert curation. A high sensitivity is achieved if the identified sets of candidate causal genes include the "true positive" causal genes at the molecular QTLs, and a high specificity is obtained if the number of identified genes that do not match the "true positive" causal genes is low. Overall, we found that there was very little variation in either of these two metrics (**Supplementary Figure S1**), indicating that the general performance of ProGeM is robust to parameter changes.

Using the default parameters, we applied ProGeM to the set of 213 mQTLs for the purposes of performance benchmarking (see below). This analysis took approximately 15 minutes using a standard Windows 7 desktop equipped with an Intel® core™ i3-3240 (3.4GHz) processor and 4GB RAM. The bottom-up, top-down, and concurrent ProGeM outputs for this set of mQTLs can be found in **Supplementary Tables S5, S6,** and **S7**, respectively.

Application and benchmarking of ProGeM

Local benchmarking. To illustrate specific characteristics of the ProGeM output in more detail, we arbitrarily selected the first (#1: rs1801133), middle (#107: rs1005390), and last (#213: rs766420) sentinel variants (**Table 1**) from the full list of 213 mQTLs (**Supplementary Table S1**), which were ordered on the basis of ascending genomic location (i.e., chr:position).

Example 1: rs1801133 and homocysteine. rs1801133 has been previously identified to associate with plasma homocysteine levels by two large-scale GWAS at genome-wide significance (32,33). rs1801133 is a missense variant (Ala222Val) affecting the *MTHFR* gene, which encodes an enzyme known to be involved in folate and homocysteine metabolism (34). Therefore, *MTHFR* was assigned to this mQTL as the expert-curated causal gene (**Table 1**). Accordingly, the ProGeM output identified *MTHFR* as the sole concurrent candidate causal gene at this mQTL, thereby highlighting the expert-curated gene as the most likely causal gene. Indeed, the bottom-up component of ProGeM indicated that *MTHFR* is the nearest protein-coding gene to rs1801133 as well as the only gene overlapping the LD region, though it was not highlighted as a *cis*-eQTL gene. Further, *MTHFR* was implicated by four out of the five top-down databases probed – the Mouse Genome Informatics (MGI) database being the exception.

Example 2: rs1005390 and X-03056—N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide. In a GWAS investigating 529 blood metabolites, rs1005390 was found to be significantly (p=2.61×10^{-16}) associated with circulating X-03056—N-[3-(2-Oxopyrrolidin-1-yl)propyl]acetamide levels (16). This variant is intronic to the expert-curated causal gene, *AOC1* (**Table 1**), which encodes an enzyme that catalyses the deamination of N1-acetylspermidine to produce the above-mentioned metabolite (35). As was the case with example 1, the expert-curated gene was identified by ProGeM as a concurrent candidate causal gene, although in this case, *AOC1* was accompanied by a second concurrent gene, *CHPF2*. Of the two genes, *AOC1* was highlighted by all three of the bottom-up data sources as well as three top-down databases. Furthermore, *AOC1* was annotated as the nearest protein-coding gene to rs1005390 whereas *CHPF2* resides distally almost 400kb away, which suggests that the nearest gene information may be useful when prioritising between concurrent candidates.

Example 3: rs766420 and bilirubin. In a GWAS investigating the genetic determinants of circulating levels of bilirubin, which is a byproduct of the breakdown of hemoglobin in red blood cells, the authors reported a significant (p=9.40×10^{-9}) association with rs766420 (36). This variant resides within an intron of the gene *TKTL1*. In contrast to the other two examples above, the expert-curated causal gene, in this case, was not the most proximal gene but rather the gene *G6PD* (**Table 1**), which is located more than 200kb downstream. *G6PD* encodes glucose-6-phosphate dehydrogenase, an enzyme that is critical for red blood cell metabolism, as deficiency is known to result in bilirubin accumulation and haemolytic anaemia (37). Although the ProGeM output for this mQTL highlighted three genes as concurrent candidates (i.e., *DNASE1L1*, *RPL10*, and *TKTL1*), none of them corresponded to the expert-curated gene, *G6PD*. *G6PD* was not highlighted by the bottom-up component of ProGeM; however, it did feature as a top-down candidate with supporting evidence from three databases. Taken together, this example underscores the importance of incorporating top-down information within the ProGeM framework, whilst also cautioning against automatically discounting non-concurrent genes without due diligence.

Global benchmarking. Following the characterisation of the ProGeM output at individual molecular QTLs, we next determined key performance indicators. In order to benchmark the sensitivity and specificity of ProGeM, we systematically compared the ProGeM output (stratified for the bottom-up and top-down component) for all 213 mQTLs with the corresponding “gold-standard” expert-curated causal gene assignments as described above (**Material and Methods**).

To assess the sensitivity, we determined the proportion of expert-curated candidate causal genes at the mQTLs (“true positives”) that were identified by ProGeM. Overall, ProGeM was able to identify the curated candidates for 209 of 213 mQTLs, thereby demonstrating high sensitivity (98%)

(**Figure 2A**). Importantly, the vast majority of these genes (*n*=174; 83%) were identified by both the bottom-up and top-down components (**Figure 2B**), indicating that sensitivity remains high even when restricting to the narrower set of concurrent candidates. Indeed, the bottom-up (**Supplementary Figure S2**) and top-down (**Supplementary Figure S3**) components alone identified 203 (95%) and 180 (85%) true candidate causal genes, respectively.

We also assessed the specificity of ProGeM across all 213 mQTLs. In total, ProGeM highlighted 1475 candidate causal genes at these 213 loci, with a median of 5 [min=2, max=30] candidates per locus (**Figure 2C**). The overall specificity of ProGeM for this dataset was 0.547 [i.e., 1530/(1530+1266)] (**Material and Methods**). When we compared the bottom-up and top-down components of ProGeM we found they performed comparably, with a specificity of 0.404 and 0.446, respectively. Indeed, the two components also had very similar medians and ranges (**Figure 2C**; **Supplementary Figure S4**). However, when only the concurrent candidate causal genes were taken into account, specificity was much improved with an overall specificity of 0.851, corresponding to a median of just 1 [min=0, max=9] candidate at each locus (**Figure 2C**).

Comparative analysis of functional annotation data sources

One of the main limitations of current bioinformatic tools for prioritising candidate causal genes at molecular QTL, including ProGeM, pertains to the difficulties associated with distinguishing between true and false positive candidates. Our benchmarking analyses showed that focusing solely on the concurrent gene set returned by ProGeM appears to be a potential means to address this issue. To formally test this, we performed an enrichment analysis using all non-concurrent candidates highlighted by ProGeM for the same mQTL data (*n*=213 loci) as a background set. A Fisher's exact test revealed a highly significant odds ratio (OR) of 28.24 ($p=1.95\times10^{-86}$), indicating that the odds of identifying the true causal genes from candidates highlighted by ProGeM are vastly improved when picking from the pool of concurrent candidate genes. Indeed, based on the observed frequencies, 48% of the concurrent candidates corresponded to an expert-curated causal gene, relative to only 3% of non-concurrent candidate genes (**Supplementary Table S8**).

Next, we assessed the importance of the various functional data sources leveraged by ProGeM for pinpointing the true positive candidate causal genes. Of all the bottom-up annotation criteria tested, the set of genes nearest to a sentinel variant was enriched with true positive candidate causal genes at the highest significance level (OR=48.19, $p=2.14\times10^{-103}$) (**Figure 3**; **Supplementary Table S8**). This was followed by the concurrent gene set (see above), the LD overlapping gene set (OR=15.27, $p=9.40\times10^{-62}$), and then the three nearest genes to a sentinel variant (OR=15.39, $p=3.89\times10^{-51}$) (**Figure 3**; **Supplementary Table S8**). Given that three of the four most highly significant criteria relate to the most proximal genes to the sentinel variant at the mQTLs, it can be concluded that proximity-based criteria are effective indicators of true positive causal genes.

When we compared the concurrent and nearest gene sets more closely, the concurrent gene set achieved higher sensitivity having identified 174 (82%), relative to 150 (70%), true positive causal genes out of 213. However, the concurrent gene set exhibited lower specificity (0.851) than the nearest gene set, which inherently achieved maximal specificity. Given this, we constructed a hybrid gene set by imposing maximal specificity onto our concurrent gene set using nearest gene information as follows: (i) for mQTLs assigned to more than one concurrent candidate gene, we restricted this assignment to the concurrent gene nearest to the sentinel variant, and (ii) we assigned mQTLs without a concurrent candidate to the gene nearest to the sentinel variant. The resultant set of

"nearest-concurrent" genes identified 163 of 213 (77%) expert-curated causal genes, which constitutes a 5% drop in sensitivity relative to our original set of concurrent genes (174; 82%) but a 7% increase relative to the nearest gene set (150; 70%). Further, an enrichment analysis of the nearest-concurrent gene set revealed both a higher odds ratio and *p*-value (OR=85.46, $p=8.33\times10^{-127}$) relative to both the original gene sets.

The annotation criterion with by far the highest odds ratio (OR=178.8, $p=5.03\times10^{-22}$) was related to genes containing a sentinel variant of moderate impact (**Figure 3**; **Supplementary Table S8**). Moderate impact genetic variants are defined as non-disruptive variants that might change protein effectiveness, which includes (i) inframe insertions and deletions, (ii) missense variants, and (iii) protein-altering variants. Indeed, of 27 such candidate causal genes highlighted by ProGeM, 26 (96.3%) corresponded to an expert-curated causal candidate (**Supplementary Table S8**). However, although genes containing a proxy ($r^2>0.8$) variant of moderate impact were also enriched with true positive causal genes, this gene set was associated with a much lower odds ratio (OR=7.67, $p=5.23\times10^{-12}$) (**Figure 3**; **Supplementary Table S8**).

Because genes containing a sentinel variant of moderate impact will also inherently be the nearest gene to that sentinel, we sought to determine whether the highly significant enrichment observed for the nearest gene set was driven by the genes that contain a sentinel variant of moderate impact. Therefore, we repeated the enrichment analysis of the nearest gene set after removing all mQTLs tagged by a moderate impact sentinel variant (note, there were no mQTLs tagged by a high impact sentinel in this dataset), which left a dataset that comprised 186 out of 213 mQTLs. We found that this filtered nearest gene set was still highly significantly enriched with true positive causal genes (OR=34.72, $p=3.80\times10^{-78}$) (**Supplementary Table S9**), indicating that the enrichment observed for the full nearest gene set (OR=48.19, $p=2.14\times10^{-103}$) was not wholly driven by the genes containing moderate impact sentinel variants. Accordingly, we obtained comparable results when we also removed mQTLs tagged by a low impact sentinel variant (i.e., synonymous, untranslated region) as well as after removing mQTLs tagged by proxy variants of high, moderate, and low impact (**Supplementary Table S9**).

The *cis*-eQTL gene sets were also significantly enriched with true positive causal genes, although the *p*-values and associated odds ratios were modest (**Figure 3**; **Supplementary Table S8**). When we investigated in more detail the 26 mQTLs for which the true positive causal gene contained a moderate impact sentinel variant, we found that just nine of these genes were also *cis*-eQTL genes for either the sentinel or a proxy variant. This indicates that these two methods of identifying true positive causal genes are predominantly exclusive. We also performed enrichment analyses for all GTEx tissues (n=44) assayed individually; however, we did not identify any specific tissues of particular relevance (**Supplementary Figure S5; Supplementary Table S10**). The same applied to individual top-down annotation criteria tested (**Figure 3**; **Supplementary Table S8**).

Cross-validation and partial replication in a large-scale pQTL dataset

Next, we assessed the extent to which these functional annotation criteria might serve as indicators of true positive candidate causal genes in a dataset from another 'omic modality. To this end, we utilised a curated *cis*-pQTL dataset comprising 561 sentinel variants, for which we hypothesised that the true positive causal gene ought to be the gene that encodes the *cis*-affected protein (**Material and Methods**; **Supplementary Table S2**). Importantly, this dataset only enabled us to assess the bottom-up annotation criteria, as the top-down criteria are not directly comparable

across the mQTL and pQTL datasets. Therefore, we used only the bottom-up candidate genes from the corresponding ProGeM output (run using default settings) as our background gene sets. Likewise, for the purposes of this comparison, we also re-ran the enrichment analyses of the bottom-up criteria for the mQTL dataset, where we used only the bottom-up candidates highlighted by ProGeM as a background gene set.

The findings for both the mQTL (**Supplementary Table S11**) and *cis*-pQTL (**Supplementary Table S12**) datasets (**Supplementary Figure S6**) were strikingly similar. First, not only did the set of nearest genes achieve the highest levels of significance in both cases (*cis*-pQTL: OR=49.55, $p=3.00\times10^{-240}$ | mQTL: OR=30.83, $p=2.69\times10^{-77}$), but the other two proximity-based criteria (LD overlapping, nearest three genes) also appeared in the top four of their respective lists (ranked by *p*-value) (**Supplementary Figure S6**). This is consistent with previous observations made for this *cis*-pQTL dataset, whereby the sentinel variants were found to cluster very closely at the transcription start site (TSS) of genes encoding the *cis*-proteins (BioRxiv: doi: https://doi.org/10.1101/134551). Second, the gene sets containing a sentinel variant of moderate impact achieved by far the highest odds ratios for both datasets (*cis*-pQTL: OR=74.64, $p=6.26\times10^{-55}$ | mQTL: OR=110.02, $p=1.81\times10^{-17}$) (**Supplementary Figure S6**). Third, the significant enrichment observed within the nearest gene sets from either dataset was only partially attenuated after removing QTLs tagged by these moderate impact sentinel variants (**Supplementary Tables S9**; **S13**). These data indicate that proximity to the sentinel variant represents a strong indicator of the true positive causal genes for both *cis*-pQTLs and mQTLs – even if the variant in question does not reside in the coding sequence.

Enrichment analyses of the *cis*-eQTL target gene sets yielded similar odds ratios across the two datasets, although higher significance was observed for the *cis*-pQTL dataset (**Supplementary Tables S11**; **12**). Further, when we performed enrichment analyses of each GTEx tissue individually, both the odds ratios and *p*-values observed for the *cis*-pQTL dataset were generally of a higher magnitude than in the mQTL dataset (**Supplementary Figures S7**; **S8**, **Supplementary Tables S10**; **S14**). However, this difference in *p*-values may be due to the fact that the *cis*-pQTL dataset comprised more than twice the number of QTLs relative to the mQTL dataset (i.e., 561 vs. 213 QTLs), thereby leveraging greater statistical power. The tissue that showed the strongest enrichment in the *cis*-pQTL dataset was the hypothalamus, whereby the *cis*-eQTL target genes derived from this brain region were significantly ($p=1.67\times10^{-7}$) enriched with true positive causal genes with an odds ratio of 41.94 (**Figure S8**; **Supplementary Table S14**).

Taken together, these results suggest that many of the same bottom-up criteria can be used to effectively prioritise the most likely true positive causal genes for both mQTLs and pQTLs, and that ProGeM may be applicable to other molecular QTL datasets beyond those tested here.

Causal gene prioritisation at loci associated with complex human diseases

Finally, we explored whether the same bottom-up criteria that applied to molecular QTLs may also be useful for prioritising the most likely causal genes at loci associated with complex human diseases. The vast majority of the sentinel variants at these loci thus far identified have been found to map to non-protein coding regions, where they may be involved in the regulation of gene expression (38,39). Therefore, proximity to a sentinel variant is generally considered to be a poor indicator of a true positive causal gene.

To the best of our knowledge, there are currently no expert-curated causal gene datasets available for complex human diseases that are sufficiently large to serve as a benchmark for comparative analyses. Thus, a comprehensive assessment of these bottom-up criteria is not yet possible in the context of complex human disease. However, for an exploratory analysis, we extracted a dataset comprising biologically plausible causal genes for coronary artery disease (3,27). After filtering, the data set comprised 56 independent loci, each assigned to a single biologically plausible causal gene (**Material and Methods**; **Supplementary Table S3**). Some of these genes have been validated experimentally and represent therapeutic targets.

We first determined how many of these curated causal genes were also the nearest to the sentinel variant tagging each respective locus. Out of 56 genes in total, 48 (85.71%) of them were found to be the nearest genes, and a further four (7.14%) of them was the second nearest. Further, we identified *cis*-eQTL targets of these 56 sentinel variants, as well as any proxies with r^2>0.8, using the GTEx data (*n*=44 tissues). We found that 23 (43.40%) of the 56 curated causal genes were *cis*-eQTL targets of either the corresponding sentinel variant or a relevant proxy. These data suggest that proximity to the sentinel variant may be a better indicator of a true positive causal gene at CAD GWAS loci than *cis*-eQTL data (based on GTEx data). Taken together, our data also suggest that the gene located closest to a sentinel variant at a GWAS locus associated with either molecular quantitative traits or complex diseases is most likely causal.

DISCUSSION

In the present study, we introduce an intuitive framework and highly configurable analysis script for the **Pr**ioritisation **o**f candidate causal **Ge**nes at **M**olecular QTLs (ProGeM). In benchmarking analyses using a set of 213 mQTLs with pre-assigned, expert-curated candidate causal genes, we demonstrated that ProGeM is highly sensitive in identifying these candidates. Further, in enrichment analyses using this mQTL dataset, we found that proximity-based indicators are an effective means of distinguishing between true and false positive causal genes.

Unique features of ProGeM. First, currently available tools are typically either trait agnostic or aimed more generally at complex disease GWAS data, whereas ProGeM is aimed at a specific trait class, i.e., molecular QTLs. This confers a major advantage, as knowledge of the trait inherently enables the incorporation of trait-specific top-down annotation criteria. In the present study, we demonstrated this for “metabolism” as a broad trait class, whereby we extracted metabolic-related genes from five freely available databases (**Material and Methods**) and cross-referenced them against genes in proximity to each corresponding mQTL sentinel variant. We note that there are other web-tools available that utilise top-down information; however, these tools are distinct in that they rely on either user-input (i.e., Phenolyzer (40)) or literature/text mining (i.e., PolySearch (41), MimMiner (42), Bitola (43), aGeneApart (44), GeneProspector (45)) to generate this information.

Second, to the best of our knowledge, ProGeM is the first available tool that integrates both bottom-up and top-down functional annotation data sources to hone in specifically on “concurrent” candidate causal genes. In our benchmarking tests, we demonstrated that the concurrent candidate causal genes for a set of 213 mQTLs were heavily enriched with the expert-curated causal candidates. These data suggest that there is a benefit in prioritising concurrent over non-concurrent candidates. Furthermore, this benchmarking also revealed multiple cases for which the expert-curated candidate

was highlighted by either bottom-up or top-down evidence, which indicates that the incorporation of two complementary layers of information is an effective means of maximising the sensitivity of ProGeM. Indeed, we observed this scenario in the example of the rs766420-bilirubin QTL discussed above, for which the expert-curated causal gene, *G6PD*, which resides ~200kb from the corresponding sentinel variant, was prioritised by the top-down component of ProGeM but not the bottom-up component.

Third, ProGeM is the first such tool to have been benchmarked against a large reference set of mQTLs with "gold standard" causal gene assignments, which allowed us to empirically verify the validity and performance of our framework. By contrast, the published tools that have been benchmarked (e.g., ToppGene Suite) have typically used very small datasets of known causal genes, which may potentially be unrepresentative (46,47). Further, benchmarking of the majority of available tools has not incorporated a baseline to empirically assess performance, but rather has comprised the application of the tool to a complex disease GWAS dataset followed by a description of the output (e.g., FUMA (21)). This lack of comprehensive benchmarking within the context of candidate causal gene prioritisation tools is due in large part to the fact that appropriate baselines have been difficult to generate. However, now that the scientific community has begun to assimilate the vast wealth of information generated by the first waves of large-scale molecular QTL studies (14,16,48,49); appropriate and large datasets are beginning to emerge. Indeed, we have made our reference mQTL dataset available in **Table S1**, providing a substrate for future benchmarking analyses and methods development.

Global benchmarking analyses. Our benchmarking analyses highlighted increased sensitivity to be one of the strengths of ProGeM, having missed only four out of 213 expert-curated causal genes at the mQTLs (**Figure 2**). When we investigated these four elusive genes in more detail, we found that three of them were missed because they are located >500kb from their respective sentinel variants (rs7542172; *AKR1A1* | rs140348140; *GLDC* | rs10403668; *LDLR*), whilst the fourth was missed because it was annotated as a pseudogene (rs7130284; *FOLH1B*). It is worth noting that rs10403668, which was associated with circulating LDL with $p<5\times10^{-8}$, is likely on the shoulder of a signal for the same metabolite tagged by another local sentinel variant, rs6511720 ($p=4\times10^{-262}$) (50). Therefore, it is possible that rs10403668 does not tag an independent mQTL, though it was retained as a sentinel variant in the present study because it fulfilled our pre-set criteria: (i) the variant resides more than 500kb from rs6511720, and (ii) the variant met the threshold for genome-wide significance (**Material and Methods**). This potentially explains why ProGeM was unable to capture the curated causal gene for rs10403668. Nevertheless, we were able to achieve maximal sensitivity by modifying two user-defined settings in ProGeM as follows: (i) omit the default filter on protein-coding genes only, and (ii) extend the genomic locus from the default +/-500kb to +/-1mb.

It is important to note that the high sensitivity achieved by ProGeM was inevitably accompanied by high levels of background noise (i.e., low specificity; **Figure 2**). This is not unusual within the context of candidate causal gene prioritisation tools, whereby the general focus tends to be on "prioritising" multiple candidates at a locus rather than force-assigning each QTL to a single causal candidate – notwithstanding that at some molecular QTLs, there may be more than one causal gene. Thus, in order to minimise the background noise associated with candidate causal gene prioritisation tools, there is a general need to be able to apply additional prioritisation strategies post-hoc that are both reliable and empirically validated. For example, as mentioned above, our comprehensive benchmarking analyses demonstrated that by specifically prioritising the concurrent candidate causal

genes highlighted by ProGeM, we were able to make considerable specificity gains at the cost of only a minimal reduction in sensitivity.

In addition to concurrent genes, our analyses also showed that proximity to the sentinel variants tagging mQTLs was another highly effective means of prioritising candidates. Indeed, out of multiple candidate gene sets defined by a series of bottom-up and top-down criteria, the set of genes nearest to each sentinel variant achieved the highest level of significance in enrichment testing, with the concurrent gene set ranking second (**Figure 3**). Furthermore, the subclass of nearest genes that contained a moderate impact sentinel variant achieved by far the highest odds ratio of all criteria tested, whereby out of 27 such genes in total, 26 corresponded to a true positive causal gene. Although this corresponds to a relatively small number of genes from the full dataset of 213 causal genes, it suggests that moderate impact sentinel-containing genes, whenever present, are remarkably reliable indicators of true causal genes for mQTLs. Importantly, we also showed that although the enrichment signal observed for the full set of nearest genes was enhanced by the subset of moderate impact sentinel-containing genes, they did not drive this signal. Thus, even in the absence of moderate-impact sentinel-containing genes, the nearest gene set still served as a reliable indicator of true positive causal genes for mQTLs.

Notably, the above findings were also recapitulated in a large *cis*-pQTL dataset where the true positive causal genes were annotated as those encoding the corresponding *cis*-affected proteins. This suggests that proximity to the sentinel variant is not only an effective means of prioritising true positive causal genes for mQTLs but also pQTLs. In contrast, for both of these molecular QTL modalities, *cis*-eQTL target genes identified in GTEx data were found to be relatively poor indicators of true positive causal genes. Nevertheless, when we extracted the *cis*-eQTL target genes for each tissue assayed by GTEx individually, the hypothalamus stood out as a potentially interesting tissue for the *cis*-pQTL but not the mQTL dataset. The hypothalamus is a major nexus point between the central nervous system (CNS) and the periphery, and is perhaps best known for its central role in the hypothalamic-pituitary-adrenal (HPA) axis, which regulates the secretion of hormones directly into the circulatory system (51). It is therefore plausible that hypothalamic *cis*-eQTLs may be detectable at the protein level in plasma, though follow-up research is required to explore this possibility.

Application to GWAS of complex diseases. In recent years, the general consensus within the context of GWAS for complex human diseases is that the underlying genetic risk factors are primarily regulatory in nature, and that the nearest gene to a sentinel variant often does not correspond to the true causal gene (38,52). In contrast, we found that a high proportion of nearest genes do correspond to the true causal genes at molecular QTLs. This suggests that the genetic mechanisms underlying complex human disease and molecular QTLs may be distinct, which is arguably logical given that disease phenotypes are invariably further away from genetics than molecular QTLs. However, in the present study we conducted preliminary further analyses using a dataset of 56 independent loci for CAD assigned to 56 curated causal genes, indicating that in most cases (85.7%) the nearest genes represented the biologically plausible candidate gene, whilst also corresponding to twice as many of the curated causal genes relative to *cis*-eQTL targets. To put this in context, for the mQTL dataset, 70.4% of nearest genes corresponded to the expert-curated genes, whereas for the *cis*-pQTL dataset the proportion was 76.3%. Thus, the nearest gene approach performed slightly better for the CAD dataset. However, we acknowledge that given the preliminary nature of this curated dataset, it is likely that a CAD risk variant located within an obviously plausible biological candidate is more likely to be assigned to a causal gene than those residing further away from their respective causal genes.

Current limitations of ProGeM. Although the curated datasets leveraged in the present study are critical for the empirical assessment of gene prioritisation tools such as ProGeM, they are not without their limitations. For example, while some molecular QTL-gene assignments have been experimentally validated, others have been supported by indirect evidence such as literature and reference databases complemented with (subjective) expert curation. Therefore, at a subset of loci, the same data sources may have been leveraged by both ProGeM and expert-curators of the mQTL and CAD datasets to assign a plausible candidate gene, thereby potentially inflating some of the observations made here. This issue does not apply to the *cis*-pQTL curation, though this dataset does come with its own potential biases. For example, plasma protein levels were quantified using nucleic acid-based aptamers (BioRxiv: doi: https://doi.org/10.1101/134551), each of which binds to a specific epitope on a target protein. As a result, the binding of these aptamers to their targets may be impacted by the presence of protein-altering variants (i.e., high or moderate impact variants) that overlap with the corresponding epitopes. Again, this may have inflated some of our observations with regards to the *cis*-pQTL dataset, though we did show that the nearest gene set was significantly enriched with true positive causal genes even after removing all *cis*-pQTLs with either a high or moderate impact sentinel variant. Nevertheless, it is reassuring that despite the different biases inherent to each of these curated datasets, our findings for each of them were very similar.

Possible extensions of ProGeM. Looking ahead, we anticipate applying ProGeM to molecular QTL datasets from additional ‘omic modalities in the future. This may be straightforward for lipid QTLs due to the similarities of standardised assay platforms but less so for *trans*-pQTLs, for example. Indeed, the prioritisation of candidate causal genes at *trans*-pQTLs would call for a different top-down strategy to the one adopted here. Accordingly, we have previously applied a “guilt-by-association” (GbA) strategy towards the annotation of *trans*-pQTLs (BioRxiv: doi: https://doi.org/10.1101/134551). Thus, rather than asking whether genes local to a sentinel variant have previously been implicated in a metabolic-related phenotype, we asked whether any local genes exhibit related functioning to the gene encoding a given trans-affected protein; i.e., annotation within the same biological pathway, or evidence of a protein-protein interaction (PPI). Notably, multiple currently available tools geared towards complex human disease have also adopted GbA strategies, which work under the assumption that unknown or novel causal genes can be identified on the basis that they must exhibit related functionality to known causal genes (47,53). There is, therefore, ample precedence for GbA within the context of candidate causal gene prioritisation.

Conclusions. In summary, ProGeM is the first gene prioritisation tool aimed specifically at identifying and prioritising candidate causal genes at molecular QTLs. We have demonstrated its utility within the context of mQTLs, with one of its major strengths being high sensitivity. We have also highlighted multiple criteria that can be used to prioritise certain candidates over others at a given mQTL. Within the ProGeM framework, those candidate causal genes with both bottom-up and top-down supporting evidence (i.e., concurrent candidates) represent likely true causal genes. We also showed that proximity to the sentinel variant is an excellent indicator of a true positive causal gene, particularly those genes containing a moderate impact sentinel (i.e., missense variants). Based on our findings, we caution against an overreliance on *cis*-eQTL target genes, as it appears that long-range regulatory effects on molecular QTLs appear may be the exception rather than the rule, and we also tentatively suggest the same may apply to complex human disease phenotypes.

AVAILABILITY

ProGeM is freely available in the form of an R script at the GitHub repository (https://github.com/ds763/ProGeM).

ACKNOWLEDGEMENT

We thank Parsa Akbari, William Astle, Jessica Barrett, James Blackshaw, Stephen Burgess, Qi Guo, Joanna Howson, Tao Jiang, Mihir Kamat, Clare Oliver-Williams, James Peters, Bram Prins, Jessica Rees and Praveen Surendran for helpful comments on the analysis framework.

FUNDING

This work was supported by the Wellcome Trust [105602/Z/14/Z to D.S.P.]. The MRC/BHF Cardiovascular Epidemiology Unit is supported by the UK Medical Research Council [MR/L003120/1]; British Heart Foundation [RG/13/13/30194]; and UK National Institute for Health Research Cambridge Biomedical Research Centre. Funding for open access charge: UK Medical Research Council [MR/L003120/1] or British Heart Foundation [RG/13/13/30194].

CONFLICT OF INTEREST

All authors declare no conflict of interest.

REFERENCES

1. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T., Farh, K.H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H. *et al.* (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421-427.
2. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, **45**, D896-D901.
3. Nelson, C.P., Goel, A., Butterworth, A.S., Kanoni, S., Webb, T.R., Marouli, E., Zeng, L., Ntalla, I., Lai, F.Y., Hopewell, J.C. *et al.* (2017) Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet*, **49**, 1385-1391.
4. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, **101**, 5-22.
5. Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M. *et al.* (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648-660.
6. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet*, **10**, 184-194.
7. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat Genet*, **39**, 1202-1207.
8. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E.T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773-777.

9. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat Genet*, **39**, 1217-1224.
10. Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J. *et al.* (2017) Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*, **49**, 131-138.
11. Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W.R., Kunze, S., Tsai, P.C., Ried, J.S., Zhang, W., Yang, Y. *et al.* (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, **541**, 81-86.
12. Lourdusamy, A., Newhouse, S., Lunnon, K., Proitsi, P., Powell, J., Hodges, A., Nelson, S.K., Stewart, A., Williams, S., Kloszewska, I. *et al.* (2012) Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet*, **21**, 3719-3726.
13. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun*, **8**, 14357.
14. Stegemann, C., Pechlaner, R., Willeit, P., Langley, S.R., Mangino, M., Mayr, U., Menni, C., Moayyeri, A., Santer, P., Rungger, G. *et al.* (2014) Lipidomics profiling and risk of cardiovascular disease in the prospective population-based Bruneck study. *Circulation*, **129**, 1821-1831.
15. Adamski, J. and Suhre, K. (2013) Metabolomics platforms for genome wide association studies--linking the genome to the metabolome. *Curr Opin Biotechnol*, **24**, 39-47.
16. Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.P. *et al.* (2014) An atlas of genetic influences on human blood metabolites. *Nat Genet*, **46**, 543-550.
17. Suhre, K., Shin, S.-Y., Petersen, A.-K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., CardioGram, Deloukas, P., Erdmann, J. *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54.
18. Brodie, A., Azaria, J.R. and Ofran, Y. (2016) How far from the SNP may the causative genes be? *Nucleic Acids Res*, **44**, 6046-6054.
19. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V. *et al.* (2015) FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med*, **373**, 895-907.
20. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. and Kastenmuller, G. (2015) SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, **31**, 1334-1336.
21. Watanabe, K., Taskesen, E., van Bochoven, A. and Posthuma, D. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*, **8**, 1826.
22. Plenge, R.M., Scolnick, E.M. and Altshuler, D. (2013) Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*, **12**, 581-594.
23. Srivenugopal, K.S. and Ali-Osman, F. (1997) Activity and distribution of the cysteine prodrug activating enzyme, 5-oxo-L-prolinase, in human normal and tumor tissues. *Cancer Lett*, **117**, 105-111.
24. Leyva-Illades, D., Chen, P., Zogzas, C.E., Hutchens, S., Mercado, J.M., Swaim, C.D., Morrisett, R.A., Bowman, A.B., Aschner, M. and Mukhopadhyay, S. (2014) SLC30A10 is a cell surface-localized manganese efflux transporter, and parkinsonism-causing mutations block its intracellular trafficking and efflux activity. *J Neurosci*, **34**, 14079-14095.
25. Moore, C., Sambrook, J., Walker, M., Tolkien, Z., Kaptoge, S., Allen, D., Mehenny, S., Mant, J., Angelantonio, E.D., Thompson, S.G. *et al.* (2014) The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials*, **15**, 363.
26. Khera, A.V. and Kathiresan, S. (2017) Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet*, **18**, 331.
27. Klarin, D., Zhu, Q.M., Emdin, C.A., Chaffin, M., Horner, S., McMillan, B.J., Leed, A., Weale, M.E., Spencer, C.C.A., Aguet, F. *et al.* (2017) Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet*, **49**, 1392.
28. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J. *et al.* (2016) PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*, **32**, 3207-3209.
29. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol*, **17**, 122.

30. Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol*, **9**, e1003118.
31. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res*, **44**, D710-D716.
32. Pare, G., Chasman, D.I., Parker, A.N., Zee, R.R., Malarstig, A., Seedorf, U., Collins, R., Watkins, H., Hamsten, A., Miletich, J.P. *et al.* (2009) Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. *Circ Cardiovasc Genet*, **2**, 142-150.
33. van Meurs, J.B., Pare, G., Schwartz, S.M., Hazra, A., Tanaka, T., Vermeulen, S.H., Cotlarciuc, I., Yuan, X., Malarstig, A., Bandinelli, S. *et al.* (2013) Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. *Am J Clin Nutr*, **98**, 668-676.
34. Mudd, S.H., Uhlendorf, B.W., Freeman, J.M., Finkelstein, J.D. and Shih, V.E. (1972) Homocystinuria associated with decreased methylenetetrahydrofolate reductase activity. *Biochem Biophys Res Commun*, **46**, 905-912.
35. Elmore, B.O., Bollinger, J.A. and Dooley, D.M. (2002) Human kidney diamine oxidase: heterologous expression, purification, and characterization. *J Biol Inorg Chem*, **7**, 565-579.
36. Sanna, S., Busonero, F., Maschio, A., McArdle, P.F., Usala, G., Dei, M., Lai, S., Mulas, A., Piras, M.G., Perseu, L. *et al.* (2009) Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum Mol Genet*, **18**, 2711-2718.
37. Kaplan, M., Hammerman, C., Vreman, H.J., Stevenson, D.K. and Beutler, E. (2001) Acute hemolysis and severe neonatal hyperbilirubinemia in glucose-6-phosphate dehydrogenase-deficient heterozygotes. *J Pediatr*, **139**, 137-140.
38. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J. *et al.* (2016) Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, **167**, 1369-1384.e1319.
39. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190-1195.
40. Yang, H., Robinson, P.N. and Wang, K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods*, **12**, 841-843.
41. Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S. and Wishart, D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res*, **36**, W399-405.
42. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. and Leunissen, J.A. (2006) A text-mining analysis of the human phenome. *Eur J Hum Genet*, **14**, 535-542.
43. Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphrey, S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, **74**, 289-298.
44. Van Vooren, S., Thienpont, B., Menten, B., Speleman, F., De Moor, B., Vermeesch, J. and Moreau, Y. (2007) Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res*, **35**, 2533-2543.
45. Yu, W., Wulf, A., Liu, T., Khoury, M.J. and Gwinn, M. (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, **9**, 528.
46. Bornigen, D., Tranchevent, L.C., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P. and Moreau, Y. (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, **28**, 3081-3088.
47. Chen, J., Bardes, E.E., Aronow, B.J. and Jegga, A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*, **37**, W305-311.
48. Del Boccio, P., Perrotti, F., Rossi, C., Cicalini, I., Di Santo, S., Zucchelli, M., Sacchetta, P., Genovesi, D. and Pieragostino, D. (2017) Serum lipidomic study reveals potential early biomarkers for predicting response to chemoradiation therapy in advanced rectal cancer: A pilot study. *Adv Radiat Oncol*, **2**, 118-124.
49. Draisma, H.H.M., Pool, R., Kobl, M., Jansen, R., Petersen, A.K., Vaarhorst, A.A.M., Yet, I., Haller, T., Demirkan, A., Esko, T. *et al.* (2015) Genome-wide association study identifies

novel genetic variants contributing to variation in blood metabolite levels. *Nat Commun*, **6**, 7208.

50. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat Genet*, **45**, 1274-1283.
51. Stacey, D., Clarke, T.K. and Schumann, G. (2009) The genetics of alcoholism. *Curr Psychiatry Rep*, **11**, 364-369.
52. Lanyi, J.K. (1986) Photochromism of halorhodopsin. cis/trans isomerization of the retinal around the 13-14 double bond. *J Biol Chem*, **261**, 14025-14030.
53. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol*, **24**, 537-544.

Table 1. Summary of the ProGeM output for three arbitrarily selected mQTLs. All bottom-up, top-down, and concurrent candidate causal genes are shown, as well as the implicating data sources. The expert-curated causal candidate genes are highlighted in red. Abbreviations: Chr; Chromosome, LD; Linkage Disequilibrium, GO; Gene Ontology, KEGG; Kyoto Encyclopedia of Genes and Genomes, MGI; Mouse Genome Informatics, Orph; Orphanet, Reac; Reactome.

Annotation					Bottom-up			Top-down						Concurrent genes
	Sentinel rsID	**Chr**	**Position** (GRCh37)	**Number proxies** ($r^2 \geq 0.8$)	**LD region overlapping genes**	**Nearest gene(s)**	***Cis*-eQTL target gene(s)**	**GO**	**KEGG**	**MGI**	**Orph**	**Reac**	**Top-scoring top-down candidate(s)**	
1	rs1801133	1	11856378	0	*MTHFR*	*MTHFR*; *C1orf167*; *CLCN6*	*MFN2*	*MTHFR*; *NPPA*; *PTCHD2*	*MTHFR*	*MAD2L2*; *NPPA*	*MTHFR*	*MTHFR*	*MTHFR* (4)	*MTHFR*
107	rs1005390	7	150543721	26	*AOC1*	*AOC1*; *TMEM179A*; *TMEM179B*	*AOC1*; *CHPF2*; *RHEB*; *TMEM179A*; *TMEM179B*	*AOC1*; *SMARCD3*	*AOC1*; *CHPF2*; *NOS3*	*NOS3*	-	*AOC1*; *CHPF2*; *NOS3*; *SMARCD3*	*AOC1* (3); *NOS3* (3)	*AOC1*; *CHPF2*
213	rs766420	X	153554404	0	*TKTL1*	*TKTL1*; *FLNA*; *TEX2B*	*BRCC3*; *DNASE1L1*; *F8A1*; *FAM3A*; *FAM50A*; *LAGE3*; *PLXNA3*; *RPL10*; *TMEM187*	*DNASE1L1*; *G6PD*; *IDH3G*; *MECP2*; *MPP1*; *RENBP*; *TKTL1*	*ATP6AP1*; *G6PD*; *IDH3G*; *TKTL1*	*DKC1*; *IKBKG*	*HCFC1*; *SSR4*; *TAZ*	*G6PD*; *IDH3G*; *RPL10*; *TAZ*	*G6PD* (3); *IDH3G* (3)	*DNASE1L1*; *RPL10*; *TKTL1*

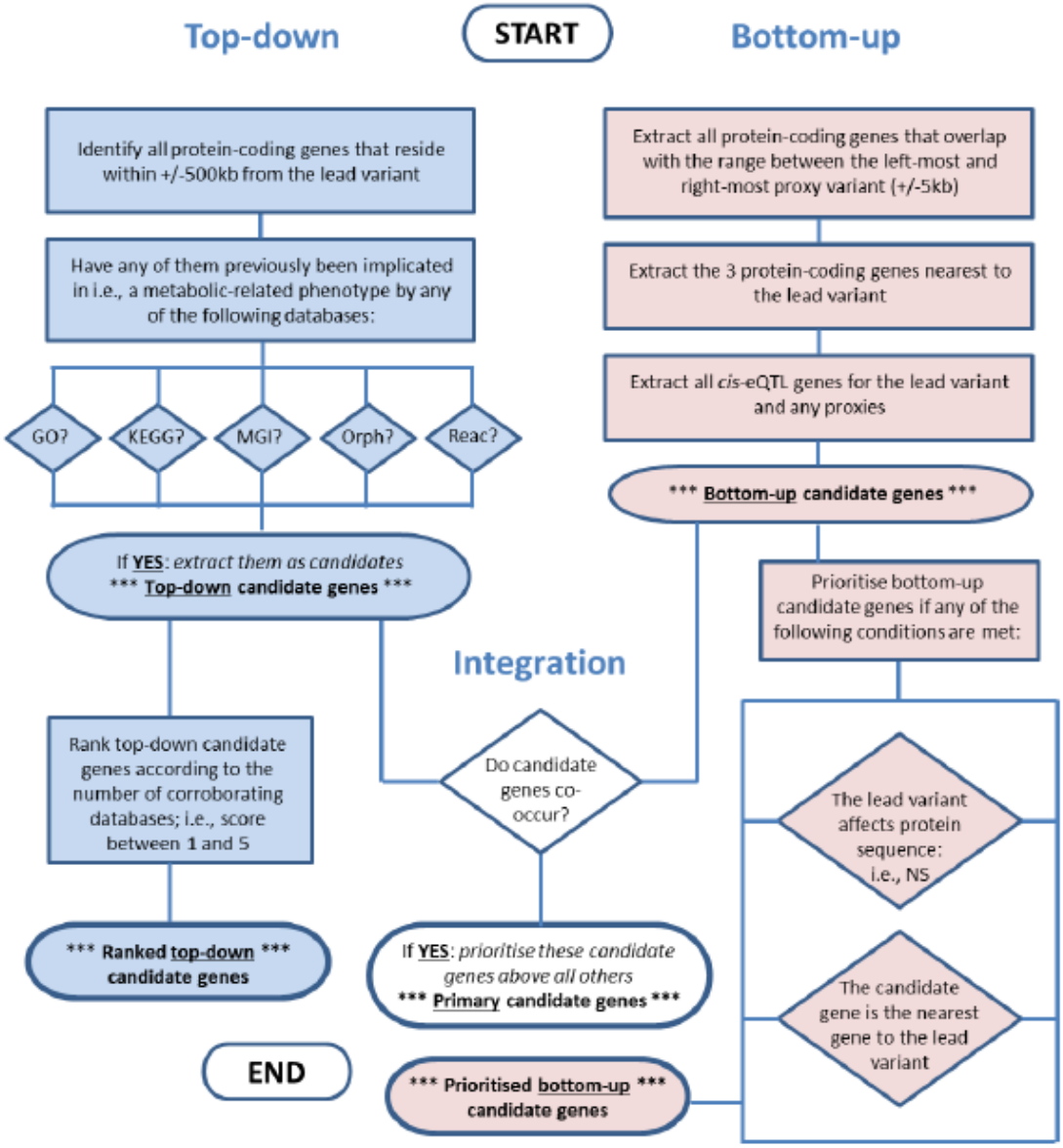

Figure 1. ProGeM: a framework for identifying and prioritising candidate causal genes at molecular QTL. Abbreviations: eQTL; expression quantitative trait locus, NS; nonsynonymous, GO; Gene Ontology, KEGG; Kyoto Encyclopedia of Genes and Genomes, MGI; Mouse Genome Informatics, Orph; Orphanet, Reac; Reactome.

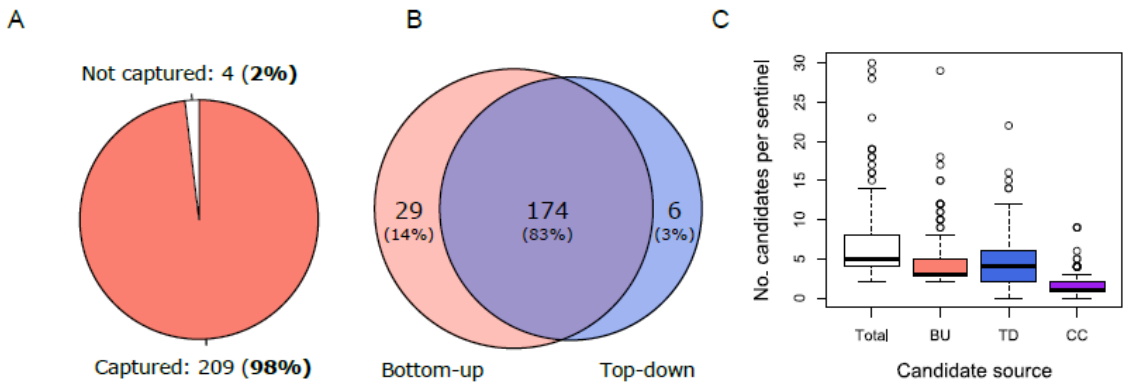

Figure 2. Global benchmarking of ProGeM using "gold standard" causal gene assignments at 213 metabolite QTLs. a) Number (percentage) of expert-curated candidate causal genes captured by our framework; b) Summary of the number of expert-curated causal genes captured either uniquely or by both the bottom-up and top-down components of our framework; c) Box plot summarising the number of candidate causal genes identified for each metabolite QTL by our framework overall (total), as well the bottom-up (BU) and top-down (TD) components uniquely and concurrently (CC). The box plot shows the median and interquartile ranges, with the whiskers extending to 1.5 times the corresponding interquartile range. Data points outside of this range are indicated individually as circles.

Data source	TP / FP	OR	adj p-value
Nearest gene	150 / 63	48.19	5.35e-102
Concurrent	174 / 189	28.24	4.89e-85
LD overlapping	169 / 274	15.27	2.35e-60
Nearest three genes	187 / 450	15.39	9.73e-50
TD score >=2	147 / 345	6.32	7.84e-31
Modifier IMPACT proxy	110 / 188	6.36	5.21e-29
Metabolic-related (Reac)	152 / 412	5.52	2.96e-26
TD score >=3	95 / 177	5.12	1.30e-21
Moderate IMPACT sentinel	26 / 1	178.80	1.26e-20
TD score >=1	180 / 701	5.00	1.60e-17
Metabolic-related (Orph)	70 / 114	5.08	3.91e-17
Sentinel cis-eQTL	105 / 282	3.52	2.34e-14
Metabolic-related (GO)	125 / 390	3.34	5.96e-14
Sentinel/proxy cis-eQTL	120 / 364	3.34	7.66e-14
Proxy cis-eQTL	110 / 337	3.06	7.97e-12
Moderate IMPACT proxy	29 / 26	7.67	1.31e-10
TD score >=4	32 / 40	5.53	3.41e-09
Metabolic-related (KEGG)	81 / 249	2.58	1.51e-07
Metabolic-related (MGI)	32 / 105	2.00	0.0695
Low IMPACT proxy	9 / 16	3.51	0.1265
TD score =5	6 / 7	5.31	0.1359
High IMPACT proxy	3 / 2	9.18	0.5632
Low IMPACT sentinel	2 / 1	12.19	1.0000
Modifier IMPACT sentinel	105 / 91	12.19	1.0000

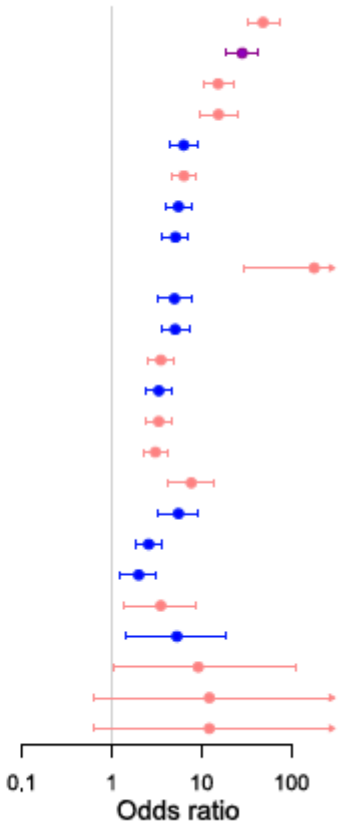

Figure 3. Comparison of top-down and bottom-up functional criteria for distinguishing true from false positive mQTL causal gene candidates. Odds ratios and 95% confidence intervals indicating the likelihood that candidate causal genes with various bottom-up or top-down characteristics correspond to the expert-curated causal genes (i.e., true causal genes). The background gene set for enrichment analysis of each characteristic was comprised of all remaining candidates identified by ProGeM. Fisher's exact test was used throughout, and Bonferroni-corrected (25 tests) *p*-values are indicated. The number of true positive (TP) and false positive (FP) causal genes identified by each characteristic are also indicated. The boxes and confidence intervals are colour-coded according to whether they correspond to bottom-up (red), top-down (blue), or concurrent (purple) data sources.