# Qtlizer: comprehensive QTL annotation of GWAS results

Matthias Munz[1-4, σ], Inken Wohlers[5, σ], Eric Simon[6], Cardiogenics Consortium, Arne S. Schaefer[4, ¶], Jeanette Erdmann[2-4, ¶, *]

[¶] These authors contributed equally

[*] Corresponding author

Email: jeanette.erdmann@uni-luebeck.de

Telephone: +49 451 3101 8301

Fax: +49 0451 3101 8304

Address:

Institute for Cardiogenetics
Universität zu Lübeck
Maria-Goeppert-Str 1
23562 Lübeck, Germany

[σ] Current address: Medical Systems Biology Group, Institute of Experimental Dermatology, Institute for Cardiogenetics, University of Lübeck, 23562 Lübeck, Germany

**Affiliations:**

[1] Institute for Cardiogenetics, University of Lübeck, 23562 Lübeck, Germany

[2] DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany

[3] University Heart Center Luebeck, 23562 Lübeck, Germany

[4] Charité – University Medicine Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute for Dental and Craniofacial Sciences, Department of Periodontology and Synoptic Dentistry, Berlin, Germany

[5] Platform for Genome Analytics, Institute of Neurogenetics and Institute for Cardiogenetics, University of Lübeck, 23562 Lübeck, Germany

[6] Target Discovery Research, Boehringer Ingelheim Pharma GmbH & Co KG, Biberach 88397, Germany

# ABSTRACT

**Motivation.** Exploration of genetic variant-to-gene relationships by quantitative trait loci (QTLs) helps to identify candidate causal variants and genes in post genome-wide association study analyses. However, the wide range of public QTL databases and the lack of batch annotation features make it cumbersome to investigate these relationships in a comprehensive manner.

**Results.** In this work, we introduce the tool 'Qtlizer' to annotate lists of common variants in human with associated changes in gene expression and protein abundance using the, to-date, most comprehensive database of published QTLs. The features include incorporation of LD variants, quality and reproducibility metrics and linking to other resources.

**Availability and Implementation.** The web application of Qtlizer is available at http://www.genehopper.de/qtlizer, a guide on how to use the REST API is available at http://www.genehopper.de/rest.

**Contact.** m.munz@uni-luebeck.de

# INTRODUCTION

In the past decade, genome-wide association studies (GWAS) led to the discovery of thousands of genetic loci that cause variation in traits and diseases. However, according to the NHGRI-EBI Catalog of published GWAS (GWAS Catalog), only a small fraction of 5-6% of the associated sentinel variants are located in protein-coding regions (**Supplementary Figure 1**). For the majority of variants the functional consequences are still unclear, resulting in a lack of mechanistic understanding of the causation. Illumination of the relationship between risk variants and genes by quantitative trait loci (QTLs) such as expression QTLs (eQTLs) can help to identify candidate causal variants and genes.

Consequently, several public databases (DBs) have evolved recently: the GTEx Portal (GTEx Consortium, 2015), Haploreg (Ward and Kellis, 2016), GRASP (Eicher *et al.*, 2015), GEUVADIS (Lappalainen *et al.*, 2013), SCAN (Gamazon *et al.*, 2010), seeQTL (Xia *et al.*, 2012), Blood eQTL Browser (Westra *et al.*, 2013), pGWAS (Suhre *et al.*, 2017), ExSNP (Yu *et al.*, 2016) and BRAINEAC (Ramasamy *et al.*, 2014), each providing specific tools and visualizations to analyze the data (**Supplementary Table 1**). However, the poor overlap of QTL datasets across the existing DBs and the lacking feature to search for multiple variants in parallel, make it cumbersome to comprehensively annotate variants in a short matter of time.

To address these shortcomings, we have developed 'Qtlizer' which extends the typical workflow of GWAS result QTL annotation by an additional step preceding the web platforms of the DBs above. Thus, we implemented a novel web-based tool on our Genehopper website which allows to explore QTL data for a given list of variants in a fast and time efficient manner by integrating all QTL datasets that were known to us and providing redirection to the source platform to allow further analyses. Moreover, we use our web application to release two complete eQTL datasets from the Cardiogenics Consortium (Garnier *et al.*, 2013; Codoni *et al.*, 2016).

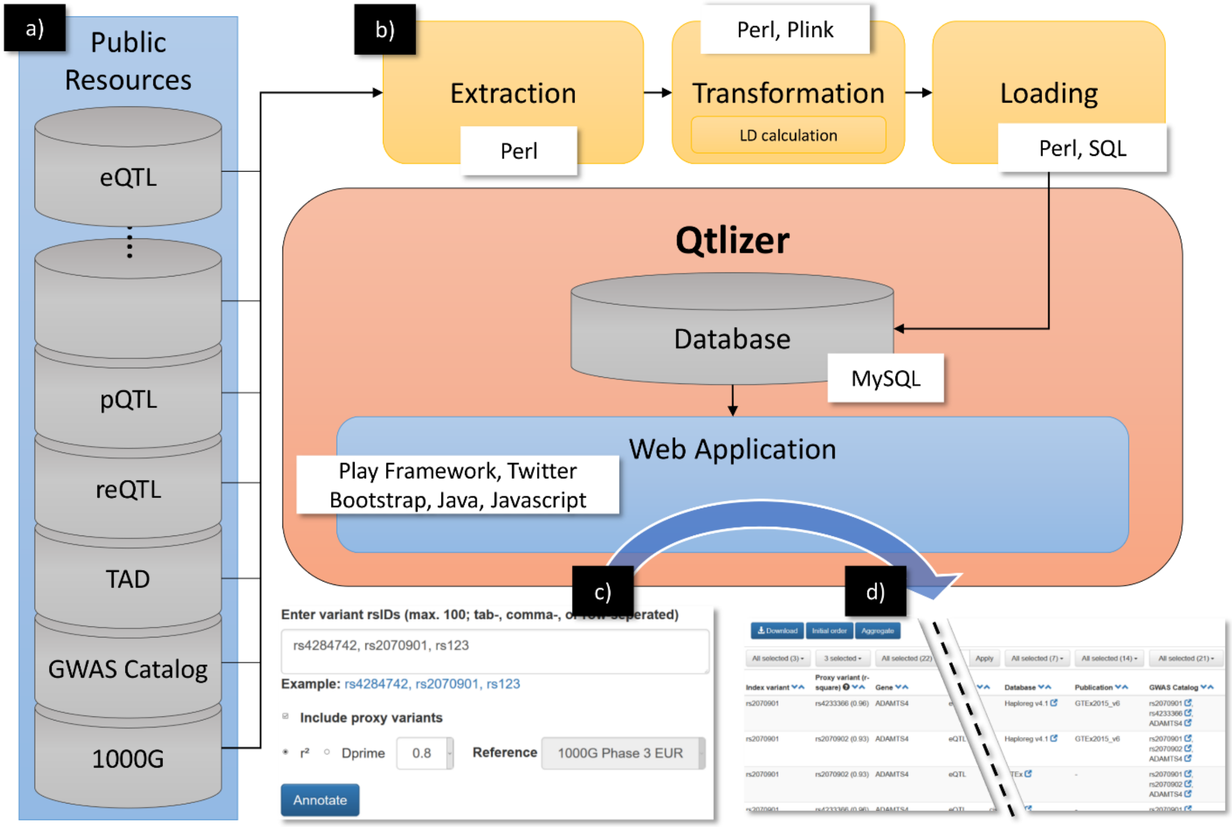
# MATERIALS AND METHODS

## Data sources

We identified 166 tissue-specific QTL datasets in the public domain and further included two inhouse datasets on monocytes and macrophages from the Cardiogenics Consortium. Of these datasets, 166 consisted of eQTLs and one each of response expression QTLs (reQTL) and protein abundance QTLs (pQTLs). A complete list of sources and datasets is shown in **Supplementary Table 1**. Additionally, we included a dataset of topological associated domain (TAD) boundaries (Way *et al.*, 2017), published variant phenotype associations from the GWAS Catalog (Welter *et al.*, 2014) and genotype data of 1000 Genomes Phase 3 (1000GP3) (Auton *et al.*, 2015). After selecting relevant datasets, we implemented an extraction, transformation and loading (ETL) process in which (a) QTL statistics were downloaded from the respective websites, (b) variants and genes were mapped to internal identifiers, tissue names were standardized and quality characteristics were calculated as well as linkage disequilibrium information based on 1000GP3 using PLINK and (c) the resulting tables were loaded into the DB of our web-based gene search engine Genehopper (available at: http://www.genehopper.de) (Munz *et al.*, 2015) (s**Figure 1. Figure 1a,b**).

## Implementation

Qtlizer was implemented in Perl (ETL process), Java (Web application backend), Javascript (Web application frontend) and SQL (DB) using the Play! Framework (available at: https://www.playframework.com) and Twitter Bootstrap (available at: https://getbootstrap.com/docs/3.3/getting-started/). We chose MySQL (https://www.mysql.com) as database management system.



Enter variant rsIDs (max. 100; tab-, comma-, or row-seperated)

rs4284742, rs2070901, rs123

Example: rs4284742, rs2070901, rs123

☑ Include proxy variants

● r² ○ Dprime 0.8 Reference 1000G Phase 3 EUR

Annotate

| Index variant | Proxy variant (r-square) | Gene | | Database | Publication | GWAS Catalog |
|---|---|---|---|---|---|---|
| All selected (3) | 3 selected | All selected (22) | | All selected (7) | All selected (14) | All selected (21) |
| rs2070901 | rs4233366 (0.96) | ADAMTS4 | e | Haploreg v4.1 | GTEx2015_v6 | rs2070901,<br>rs4233366,<br>ADAMTS4 |
| rs2070901 | rs2070902 (0.93) | ADAMTS4 | eQTL | Haploreg v4.1 | GTEx2015_v6 | rs2070901,<br>rs2070902,<br>ADAMTS4 |
| rs2070901 | rs2070902 (0.93) | ADAMTS4 | eQTL | TEx | - | rs2070901,<br>rs2070902,<br>ADAMTS4 |
| rs2070901 | rs4233366 (0.96) | ADAMTS4 | eQTL | | - | rs2070901 |

**Figure 1.** (a) Data from various public available resources was integrated (b) applying an Extraction, Transformation and Loading (ETL) process. (c) Qtlizer can be queried for QTL data via the web-based user interface or programmatically by inputting lists of genetic variants using a REST API. (d) The annotation results are displayed in a table view.

## RESULTS AND DISCUSSION

With Qtlizer we introduce a web-based tool to annotate lists of genetic variants with QTL data in a comprehensive way. After applying the ETL process described in the last section, 40,883,209 QTLs (37,014,094 study-wide significant) from 3,856,968 variants to 32,987 genes were finally added to the Genehopper DB. The user interface of Qtlizer takes a list of index variants in the form of reference SNP identifiers (rsIds) as input (s**Figure 1. Figure 1c**). Optionally, these variants can be enriched for proxy variants using the European reference population of 1000GP3 by setting a threshold for $r^2$ or D'. The resulting QTLs are displayed in a table view, where each line represents a single QTL (**Figure 1d)**. The table can be sorted and filtered column-wise and the current table state can be downloaded as text file for further use. If index variants were enriched for proxy variant, results can be aggregated to create a haplotype block centric view in which QTLs of the proxy variants and the index variant to a specific gene are shown in one line. Along with the columns for variant, gene, type of QTL, distance in base pairs, tissue, effect size, significance information, and origin, we added three more attributes: (1) Co-localization: instead of defining fixed size of e.g. one mega base pair, we utilized the TAD boundaries to categorize QTLs into *cis* (i.e. variant and gene remain in the same TAD) and *trans*, because it has been shown that regulatory DNA elements and gene promoters are interacting more frequently within a TAD whose boundaries are primarily cell-type independent (Way *et al.*, 2017). (2) Relevance flags and counts: these were added to each QTL to draw conclusions about its reproducibility across studies as well as its causality, e.g. whether it is the best QTL for a specific gene and tissue in terms of *P*-value (3) GWAS Catalog: variants and genes in the result set are tagged if they are listed in the catalog. Further, each result is linked with the corresponding publication and/or source database to allow closer investigation. We restricted the number of index variants to 100 and the number QTLs that are returned to 5000. For larger queries we provide a web service, allowing programmatic access to the QTL data (**Supplementary Figure 2, 3**).

In summary, Qtlizer provides a web-based solution for annotating lists of genetic variants with QTLs from a large number of published datasets and databases as well as from two recently released eQTL datasets generated by the Cardiogenics Consortium.

## ACKNOWLEDGMENTS

This work was financially supported by the cluster of excellence “Inflammation at Interfaces” (IaI). We thank the Cardiogenics Consortium for data related to the macrophage and monocyte eQTLs. Members of the Cardiogenics Consortium are listed in **Supplementary Table 2**.

## FUNDING SOURCES

This work was supported by a research grant of the German Research Foundation DFG (Deutsche Forschungsgemeinschaft; GZ: SCHA 1582/3-1). I.W. was supported by the Peter und Traudl Engelhorn Foundation.

# REFERENCES

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Codoni,V. *et al.* (2016) Preservation Analysis of Macrophage Gene Coexpression Between Human and Mouse Identifies PARK2 as a Genetically Controlled Master Regulator of Oxidative Phosphorylation in Humans. *G3 (Bethesda).*, **6**, 3361–3371.

Eicher,J.D. *et al.* (2015) GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, **43**, D799-804.

Gamazon,E.R. *et al.* (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–62.

Garnier,S. *et al.* (2013) Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS Genet.*, **9**, e1003240.

GTEx Consortium,Gte. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–60.

Lappalainen,T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–11.

Munz,M. *et al.* (2015) Multidimensional gene search with Genehopper. *Nucleic Acids Res.*, **43**, 98–103.

Ramasamy,A. *et al.* (2014) Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.*, **17**, 1418–1428.

Suhre,K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.*, **8**, 14357.

Ward,L.D. and Kellis,M. (2016) HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–D881.

Way,G.P. *et al.* (2017) Implicating candidate genes at GWAS signals by leveraging topologically associating domains. *Eur. J. Hum. Genet.*

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001-6.

Westra,H.-J. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, **45**, 1238–43.

Xia,K. *et al.* (2012) seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, 451–2.

Yu,C.-H. *et al.* (2016) Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease. *OMICS*, **20**, 400–14.