PLOS COMPUTATIONAL BIOLOGY

SOFTWARE

BALDR: A Web-based platform for informed comparison and prioritization of biomarker candidates for type 2 diabetes mellitus

Agnete T. Lundgaard[1], Frédéric Burdet[2], Troels Siggaard[1], David Westergaard[1], Danai Vagiaki[1], Lisa Cantwell[1], Timo Röder[1], Dorte Vistisen[3,4], Thomas Sparsø[5], Giuseppe N. Giordano[6], Mark Ibberson[2], Karina Banasik[1], Søren Brunak[1]*

1 Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, Copenhagen, Denmark, **2** Vital-IT, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, **3** Clinical Epidemiological Research, Steno Diabetes Center Copenhagen, Herlev, Denmark, **4** Department of Public Health, University of Copenhagen, Copenhagen, Denmark, **5** Bioinformatics and Data Mining, Global Research Technologies, Novo Nordisk A/S, Måløv, Denmark, **6** Genetic and Molecular Epidemiology Unit, Lund University Diabetes Centre, Department of Clinical Sciences, Clinical Research Centre, Lund University, Skåne University Hospital, Malmö, Sweden

* soren.brunak@cpr.ku.dk

OPEN ACCESS

Citation: Lundgaard AT, Burdet F, Siggaard T, Westergaard D, Vagiaki D, Cantwell L, et al. (2023) BALDR: A Web-based platform for informed comparison and prioritization of biomarker candidates for type 2 diabetes mellitus. PLoS Comput Biol 19(8): e1011403. https://doi.org/10.1371/journal.pcbi.1011403

Editor: Majid Jaberi-Douraki, Kansas State University, UNITED STATES

Received: November 21, 2022

Accepted: July 31, 2023

Published: August 17, 2023

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: Funding for the work performed has been provided within the framework of the RHAPSODY Consortium. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115881 (RHAPSODY). RHAPSODY receives support from the European Union's Horizon 2020 research and

Abstract

Novel biomarkers are key to addressing the ongoing pandemic of type 2 diabetes mellitus. While new technologies have improved the potential of identifying such biomarkers, at the same time there is an increasing need for informed prioritization to ensure efficient downstream verification. We have built BALDR, an automated pipeline for biomarker comparison and prioritization in the context of diabetes. BALDR includes protein, gene, and disease data from major public repositories, text-mining data, and human and mouse experimental data from the IMI2 RHAPSODY consortium. These data are provided as easy-to-read figures and tables enabling direct comparison of up to 20 biomarker candidates for diabetes through the public website https://baldr.cpr.ku.dk.

Introduction

The advance of high-throughput omics methods has given rise to large-scale data capture for biomarker discovery. However, with ever-increasing data density from these analyses, the ability to prioritize candidates of interest for follow-up experiments, as part of an efficient downstream verification process, has become increasingly necessary. While many online tools exist for the analysis of single molecular entities in the context of human disease, few tools are available for the comparison and prioritization of a large set of molecules as potential biomarkers [1–3]. With the rapid rise in type 2 diabetes mellitus (T2DM) incidence worldwide [4,5], there is an increasing need for new diagnostic and progression biomarkers. However, there are currently no integrative tools available for diabetes biomarker discovery or prioritization. To address this gap, we developed Biomarker AnaLysis for Diabetes Research (BALDR), a tool that automatically produces a comprehensive report that enables informed comparison of up to 20 targets for their relevance as biomarkers for T2DM.

innovation programme and EFPIA. This work is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies. Furthermore, this work were supported by the Novo Nordisk Foundation (grants NNF14CC0001 and NNF17OC0027594 to A.T.L., T.Si., D.W., D.Va., L.C., T.R., K.B., and S.B.), Bayer A/S (research grants to D.Vi.), Sanofi Aventis (research grants to D.Vi.), Novo Nordisk A/S (research grants to D.Vi.), and Boehringer Ingelheim (research grants to D. Vi.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: D.Vi. is employed at Novo Nordisk A/S and holds shares in Novo Nordisk A/S. T.Sp. is employed at Novo Nordisk A/S and holds shares in Novo Nordisk A/S. T.Si. holds shares in Armata Pharmaceuticals Inc., Intellia Therapeutics Inc., Moderna Inc., and IGM Biosciences Inc. S.B. holds shares in Intomics A/S, Hoba Therapeutics Aps, Novo Nordisk A/S, Lundbeck A/S, and managing board memberships in Proscion A/S and Intomics A/S. All other authors have no conflicts of interest to declare.

Public databases such as UniProt [6], PHAROS [7], Open Targets [8], DrugBank [9], and the Human Protein Atlas (HPA) [10] contain vast amounts of functional, interactional, and disease-relevant information for the human proteome. In BALDR, we utilize these data to inform on the relevance of proteins as novel, biologically relevant biomarkers for diabetes. Moreover, we include observational and experimental data from the Risk Assessment and ProgreSsiOn of Diabetes (RHAPSODY) consortium, financed by the EU Innovative Medicines Initiative-2. Within RHAPSODY, a large quantity of omics data has been generated based on a federation of clinical cohorts that include patients in varying stages of T2DM [11], as well as human tissue analysis [12,13] and mouse experiments [14] conducted within the consortium. These data include analysis of human blood and pancreatic islets and mouse adipose, skeletal muscle, liver, and pancreatic islet tissue, measuring a total of over 16,000 proteins and protein-coding gene transcripts.

BALDR is, to our knowledge, the first publicly available tool that allows for direct comparison between multiple protein biomarker candidates (Fig 1). We facilitate the integration and comparison of user candidates in a set of user-friendly graphics and tables that can readily be used for scientific publications. In the present version of BALDR, we use a combination of experimental data, text mining of full-length papers, and data obtained from major publicly available databases and repositories. This enables the user to make informed comparisons and prioritizations of biomarker candidates for T2DM. The framework is not limited to T2DM but can easily be adapted to other diseases of interest by changing the data capture workflow accordingly. BALDR is provided as open access through the public website (https://baldr.cpr.ku.dk).

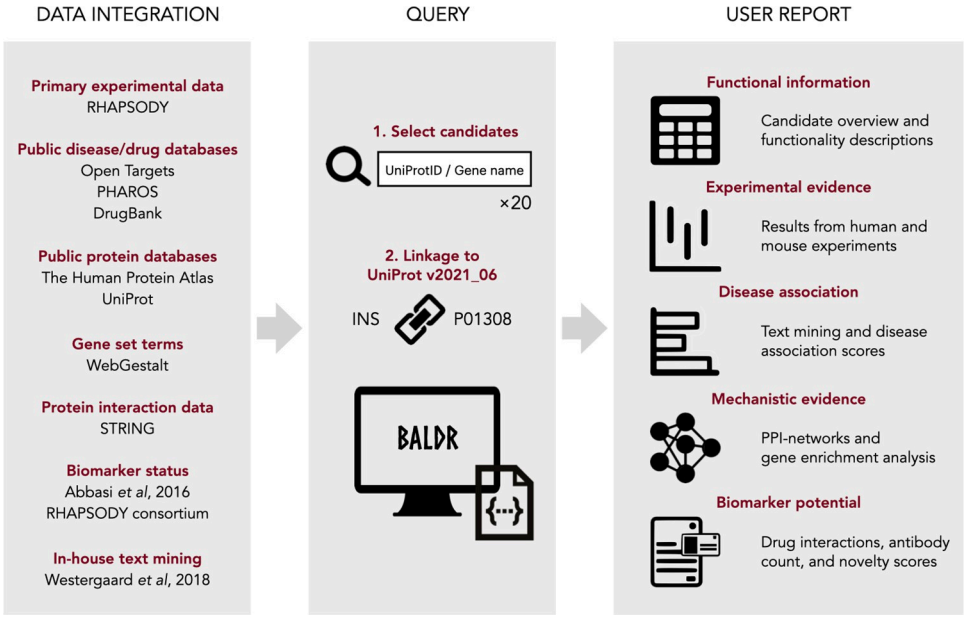

Fig 1. Schematic representation of the BALDR pipeline. The BALDR pipeline consists of three steps: Data integration, Query, and User report. In the Data integration step, data from multiple data sources is captured and consolidated into the BALDR database accessible to the BALDR source code. In the query step, the user selects up to 20 candidates using UniProt IDs or gene names. The targets are linked by UniProt IDs to the BALDR database and the BALDR source code runs using Rmarkdown. Lastly, a User report is produced containing five main sections: Functional information, Experimental evidence, Disease association, Mechanistic evidence, and Biomarker potential. Image credit: Openclipart.org.

https://doi.org/10.1371/journal.pcbi.1011403.g001

Design and implementation

Data capture for biomarker prioritization

We extracted all relevant data from five public data sources included in the BALDR framework as a snapshot at the time of compilation. Specifically, we captured data from UniProt [6] (https://www.uniprot.org/), PHAROS [7] (https://pharos.nih.gov/), Open Targets [8] (https://www.opentargets.org/), DrugBank [9] (https://www.drugbank.com/), and the Human Protein Atlas (HPA) [10] (http://www.proteinatlas.org).

The UniProt database contains protein annotation and sequence data for the human proteome, as well as multiple other species. Data from the UniProt database was used for mapping genes and proteins using the UniProt and Ensembl IDs.

The PHAROS database is developed to improve drug discovery by gathering information for poorly annotated potential drug targets, especially G-coupled receptors, ion channels, and kinases. Within PHAROS, scores have been developed to characterize the knowledge level for targets. The PHAROS database was used for protein-specific information in the form of descriptions and aggregated scores for functionality and development level.

The Open Targets database consists of data collected from multiple sources on disease-target associations and summarizes this data as evidence scores for multiple data types, including text mining, genetic associations, and animal models. An overall score across data types is also provided. Diseases and phenotypes can be searched using the Experimental Factor Ontology (EFO) and Mondo Disease Ontology (MONDO). Open Target data was used here to inform on the evidence availability for targets in relation to diabetes mellitus (EFO_000400), as well as T2DM (MONDO_0005148), T1DM (MONDO_0005147), GDM (EFO_0004593), and other diabetes types.

The DrugBank database contains information on drugs and their association with molecular entities. Here, we include information on all drugs with evidence of interaction with a protein target. We include basic information on the drug, including ATC codes (if available), drug group, and pharmacological action. We also include links to evidence for these interactions through the DrugBank platform.

The HPA database contains information on the localization of proteins in cells, tissues, and organs based on omics and imaging data. HPA data was here used to inform on protein class and cellular location. Data from HPA was further used to identify proteins found in circulation for the RHAPSODY meta-analysis (see below).

The semi-automatic data capture pipeline is based on the Snakemake workflow system [15], integrating individual Python scripts for each database. The pipeline can be adapted to any disease with an Experimental Factor Ontology (EFO) label for the capture of disease-specific data from Open Targets and PHAROS. The data from these five major protein, gene, and disease databases resulted in 70 selected unique features for 20,343 of the 20,375 proteins available in UniProt (Table 1 and S1 File).

Table 1. Database overview.

Database	Variable count	Protein count
UniProt	5	20,375
PHAROS	12	20,268
Open Targets	42	18,881
DrugBank	11	3,025
The Human Protein Atlas (HPA)	4	19,139
Total	**70**	**20,343**

https://doi.org/10.1371/journal.pcbi.1011403.t001

Suggested biomarkers for type 2 diabetes

A table of known and suggested biomarkers for diabetes was compiled using three sources, i) Abbasi *et al* [16] conducted a systematic review of biomarkers T2DM incidence risk, identifying 167 potential biomarkers for diabetes. Of these, we included 48 proteins with a UniProt ID; ii) A biomarker task force within the RHAPSODY consortium curated a list of 33 biomarker candidates; iii) A network expansion was performed on the 81 biomarkers candidates found in i) and ii) using the Intomics InBio protein-protein interaction data [17]. These were compared to a network based on the 81 diabetes-associated sites previously identified by GWAS [18]. Of these, five genes were found to be overlapping and secreted. Finally, we performed a network expansion and identified 47 interaction partners, which interacted with at least two of the 81 suggested biomarkers with a confidence of at least 0.1. Together, 133 biomarkers were included as known or suggested biomarkers for T2DM. The full list is provided in S2 File.

RHAPSODY data and meta-analysis

The RHAPSODY project aims to identify novel biomarkers for T2DM susceptibility and progression through analysis of omics data from twelve European cohorts comprising a total of 68,000 individuals with different stages of T2DM. Moreover, the project aimed to determine the link between alterations of insulin secretion and insulin action in the liver, adipose tissue, and skeletal muscle using pre-clinical mouse models. Lastly, β-cell functional study data from human and mouse pancreatic tissue are also available through the consortium. Dependent on species and tissue, multiple omics measurements have been performed, including, but not limited to, targeted metabolomics, lipidomics, genomics, transcriptomics, and proteomics.

Here, we included data from proteomics and transcriptomics experiments performed in human and mouse studies from four main analyses: Cox proportional hazard models for diabetes progression on 1,195 proteins using the SomaLogic SOMAscan platform in 1,188 individuals from the DCS and GoDARTS cohorts as described by Slieker *et al* [11]; differential RNA expression analysis in human islets and pancreatic tissues from 84 non-diabetic (ND) and 19 T2DM organ donors and 103 pancreatic tissue specimens (32 ND individuals, 36 T2DM individuals, 15 individuals with impaired glucose tolerance (IGT), and 20 with type 3c diabetes (T3D)), as described by Solimena *et al* [12] and Wigger *et al* [13]; and differential RNA expression analysis of adipose tissue, pancreatic islets, liver tissue, and skeletal muscle tissue from three mouse strains (C57Bl/6J, DBA/2J, and BALB/cJ) with high fat, high sucrose diet-induced glucose dysregulation (HF) compared to regular diet (RD)-fed animals, as described by Sanchez-Archidona *et al*[14].

We performed a sample size-based p-value meta-analysis, as described by Willer, Li, and Abecasis [19] across all experiments listed above. The resulting meta p-values were adjusted for multiple testing using false discovery rate (FDR) correction [20]. To rank the targets, we created a subset of targets that i) had an adjusted meta p-value under 0.05; ii) have been identified as found in circulation, either by expression or secretion to the bloodstream, using the Human Protein Atlas data; and iii) were included in the full-length text mining data, described below. To identify the most novel leads, the resulting targets were then inversely ranked by the number of co-mentions with diabetes so that the lowest number of co-mentions (zero) resulted in the best score of one. A total of 5035 (25%) proteins were assigned a rank between 1 and 710, with the best rank of 1 assigned to 367 targets with zero (0) co-mentions with diabetes.

Text mining

Text mining was conducted on 15 million full-text scientific articles and their corresponding abstracts using the method described in Westergaard *et al* [21] with two settings. In short, 1,488,927 articles from the open-access PubMed Central corpus (PMC), 3,335,400 articles from Springer, and 11,697,096 articles from Elsevier spanning the period 1823–2016 were collected. Of these, 902,415 articles were removed as not written in English and additional 1,069,525 articles were removed in quality control, yielding 14,549,483 full-text articles for text mining. Additional 16,544,511 abstracts from MEDLINE were included for text mining. Text mining was performed using Named Entity Recognition (NER) using a dictionary comprised of gene names from the STRING database [36] and disease names from Disease Ontology [22]. A weighted score was calculated based on weighted counts, where co-occurrence within the same sentence or paragraph gives higher counts than co-occurrence within the same document. Weighted scores were calculated for both full-text articles and abstracts in two searches. For one search, all co-mentions between "diabetes" and all proteins were pulled, resulting in 16,366 proteins with at least one co-mention out of the 18,563 proteins (88%) identified in the text corpus. Similarly, we pulled co-mentions between disease terms for all diseases *but* diabetes and proteins and found at least one co-mention for 18,212 proteins (98%). Using full-text articles yielded on average 40 times more co-mentions for diabetes and 22 times more co-mentions for all other diseases compared to text mining of abstracts.

Protein-protein interaction networks

Public protein-protein interaction (PPI) network data from the STRING database [23] were used to build PPI networks for each target. The networks were built to include physical interactions between candidates and interaction partners (primary interactions), as well as between interaction partners (secondary interactions). We used the recommended confidence threshold of 0.7 for the combined score to differentiate between high-confidence interactions and potential interactions [24]. The current version of BALDR implements the 9606 (human) physical links protein network data v11.5.

Overrepresentation analysis

WebgestaltR [25] was used to conduct an overrepresentation analysis (ORA) of functional gene set terms on eight databases (see S3 File) in two separate analyses. In the first analysis, the enrichment analysis was conducted on the protein-protein interaction networks identified in the PPI analysis against all transcripts and proteins measured in RHAPSODY. Only terms including the candidates of interest were kept. In the second analysis, the enrichment analysis was conducted on the selected candidates in a shared analysis against all transcripts and proteins measured in RHAPSODY. In both analyses, enrichment is shown as significant (FDR adjusted p-value < 0.05) and not significant (FDR adjusted p-value ≥ 0.05). Based on the PPI networks, we performed gene enrichment analysis for 15,185 networks. Of these, 12,475 (82%) had at least one enriched gene term.

User report

BALDR is built using R v4.2.1 [26] and R markdown v2.11 [27], a plain text editor that integrates R code and outputs with HTML code to produce formatted documents. The R packages ggplot2 v3.3.5, knitr v1.33 [28], and formattable v0.2.1 [29] were used to produce formatted figures and tables that are provided as stand-alone files in multiple formats. Figures are

provided both as standard .png files and in vector format as .pdf files. Tables are likewise provided as .png files, as well as Excel sheets.

The BALDR report contains five sections that cover the different aspects of the input data:

i. *Functional information*. To provide a basis for further exploration of the chosen candidates, we give a general overview of the candidates' protein characteristics in the first table. The variables include subcellular location (HPA), protein family (PHAROS), protein class (HPA), and target development level (PHAROS). A second table presents functional descriptions from PHAROS, as well as their status as diabetes biomarkers.

ii. *Experimental evidence*. Experimental data from the RHAPSODY consortium is presented through volcano plots, scatter plots, and tables. We provide comparisons between candidates, as well as omics, species, and tissue types. Moreover, we provide a meta-analysis and ranking of all targets according to text-mining novelty for diabetes. This enables the user to compare their own results to those obtained in the RHAPSODY consortium and to assess the strength of the evidence across candidates.

iii. *Disease association*. To explore existing evidence for candidate-disease associations, we present disease-wide text-mining data (PHAROS and in-house text mining), diabetes-specific text-mining data from 15 million full-text articles (in-house text mining), and association scores for diabetes mellitus as a supergroup, as well as selected diabetes mellitus types such as T2DM, T1DM, and GDM (Open Targets).

iv. *Mechanistic evidence*. We use public protein-protein interaction data to produce protein-protein interaction networks for each candidate separately. Interaction partner information is made available to enable pathway exploration. These interaction networks also serve as a basis for gene set enrichment analysis for GO-terms, pathways, and disease/drug association. In a separate analysis, the input targets are analyzed together for enriched gene set terms to explore commonalities between candidates.

v. *Biomarker potential*. In the last section, we explore the candidates in the context of their potential as biomarker candidates. We provide information on target-drug interaction from DrugBank, counts of commercially available antibodies, and a collective novelty score based on text mining from PHAROS. These data should be seen in the context of the candidates as general biomarkers and may not reflect their potential as diabetes biomarkers.

We showcase BALDR using the six novel biomarkers significantly associated with diabetes progression as measured by time to initiation of insulin treatment in the DCS and GoDARTS cohorts, as identified by Slieker *et al* [11]. The six proteins were identified as having the highest effect size for acceleration to insulin dependency out of 11 proteins significantly associated with the outcome. We compared these to three biomarkers associated with T2DM in the greatest number of incident cases included in the systematic review by Abbasi *et al* [16] (Table 2).

Code availability and access

The code for BALDR and the depending data capture is freely available on GitHub at https://github.com/agnetelundgaard/BALDR.

Results

Automated data capture workflow combining multiple data sources

Comparing proteins as potential biomarkers, in the context of a specific disease such as T2DM, is challenging as candidate ranking can be based on different features, such as novelty,

Table 2. Showcase targets from Slieker et al, 2021 [11] and Abbasi et al, 2016 [16].

Source	Gene name	Protein name	UniProt ID
[11]	CRELD1	Protein disulfide isomerase CRELD1	Q96HD1
	ENPP7	Ectonucleotide pyrophosphatase/phosphodiesterase family member 7	Q6UWV6
	FAS	Tumor necrosis factor receptor superfamily member 6	P25445
	GDF15	Growth/differentiation factor 15	Q99988
	IL18R1	Interleukin-18 receptor 1	Q13478
	RTN4R	Reticulon-4 receptor	Q9BZR6
[16]	CRP	C-reactive protein	P02741
	GGT1	Glutathione hydrolase 1 proenzyme	P19440
	GPT	Alanine aminotransferase 1	P24298

https://doi.org/10.1371/journal.pcbi.1011403.t002

protein interaction partners, or gene-disease evidence depending on the research objective. To address this challenge, we made a semi-automated pipeline that captures public data from five major protein databases (UniProt, PHAROS, Open Targets, DrugBank, and the Human Protein Atlas (HPA)) that serve as the main data input for BALDR. We further enriched the report with data generated by text-mining of 15 million full-length papers [21], public protein-protein interaction network data [23], functional gene enrichment data [25], T2DM biomarker suggestions from internal and external sources [16], and published data from the RHAPSODY consortium [11–14]. We have gathered and standardized selected features for 20,343 human proteins for which comparative figures and tables can be produced through the BALDR pipeline. The easy-to-read graphics and tables can be used in publications with minimal editing required or imported by expert users to produce new graphics. Presenting these data directly to the user allows flexible decision-making and encourages transparency of the prioritization process.

Example of a BALDR report

To showcase the value of BALDR, we compared nine potential diabetes biomarkers, six protein biomarkers identified by Slieker *et al* [11] as significantly correlated with T2DM progression in multiple cohorts, and three protein biomarkers identified by Abbasi *et al* [16] to have been positively associated with T2DM in the greatest number of incident cases (Table 2). We use this as a case scenario, where the task was to prioritize promising biomarker candidates identified in primary experimental data and we compare these to three highly studied T2DM biomarkers as a reference point. The example report on the nine biomarkers can be found in the supplement material (S4 File).

In section 1—Basic information, we find the cellular location, protein family, and PHAROS target development level of the candidates, as well as detailed functionality descriptions. ENPP7, GGT1, and GTP are enzymes involved in lipid or amino acid metabolism, while the other candidates belong to the non-IDG family, *i.e.*, proteins that are not expected to be likely drug targets. The majority of these are receptors or receptor ligands involved in cellular signaling. One target, CRELD1, does not have a functionality description from PHAROS, so here, we need to extend our search to UniProt via the provided link. The section also provides summarized information from Open Targets on the association of candidates with diabetes. All except CRELD1 and ENPP7 have previously been directly or indirectly associated with T2DM, as well as other diabetes types.

In the next two sections, section 2 –Experiment evidence and section 3 –Disease association, we found that FAS, CRELD1, GDF15, CRP, and GPT were all found to be significantly

associated with diabetes in a minimum of one RHAPSODY experiment, while only the Slieker *et al* [11] candidates were significant in the RHAPSODY cohort meta-analysis. In this meta-analysis, the highest ranked target is ENPP7, indicating that the target had the lowest count of co-mentions with diabetes, a proxy for the target's novelty as a diabetes biomarker. Interestingly, when we look at the disease association data, we can see that ENPP7 has the lowest text mining score for all metrics but the highest fraction of diabetes co-mentions to all-disease co-mentions except for CRP, which also has the highest total number of co-mentions. Similarly, in the later section on target novelty (section 5.3), ENPP7 has one of the highest novelty scores. Taken together, ENPP7 appears to be understudied in general but may be relevant as a novel biomarker for diabetes. In contrast to ENPP7, we find that CRP, FAS, and GDF15 are the most studied targets for diabetes. GDF15 was identified by Slieker *et al* [11] to have the greatest effect size in their proteomics analysis, as shown in plot 2.1.2 in the report. Moreover, CRP was identified as the target with the highest association score for T2DM, as shown in section 3.2, followed by GDF15. Conversely, CRELD1 and ENPP7 were not associated with T2DM or any other diabetes type in Open Targets (sections 1.2 and 3.2).

In section 4 –Mechanistic evidence, protein-protein interaction (PPI) networks and gene enrichment are explored. The two largest networks are found for RTN4R and FAS, containing 201 and 117 interactions, respectively. This is explained by their role in large signaling pathways, as RTN4R is a receptor in the Rho signaling pathway responsible for the reorganization of the cytoskeleton, while FAS is a receptor for Caspase signaling mediating apoptosis. In contrast, we find no interaction partners for CRELD1 and no high-confidence interaction partners for ENPP7, GGT1, and GTP. IL18R1 has four high-confidence interaction partners, all related to interleukin signaling, including IL18 identified by Abbasi *et al* [16] as a biomarker for diabetes. CRP has multiple high-confidence interaction partners which have been identified as potential T2DM biomarkers through network expansion (FCN2 and CFH), by Abbasi *et al* [16](C3) or has been suggested by experts in the RHAPSODY consortium as potential novel targets (OLR1). FCN2, CFH, and C3 are involved in innate immunity, while OLR1 is involved in the degradation of oxidized low-density lipoprotein (oxLDL). Lastly, GDF15 has two high-confidence interaction partners, RET and GFRAL, that together mediate GDF15-induced food restriction.

Looking at the gene enrichment analysis we see a relatively limited number of shared terms for the KEGG pathways, GLAD4U diseases, and GLAD4U drugs databases (section 4.2.2). Most notably, FAS, IL18R1, and GDF15 have a non-significant enrichment in inflammatory pathways and anti-infectives. This is similarly seen in the gene enrichment analysis of their networks, where terms related to inflammation and apoptosis are found in most of the searched databases. Moreover, we find CRP, GGT1, GPT, and GDF15 to be highly enriched for multiple disease terms related to cardiovascular diseases and diabetes in the GLAD4U disease database. We also find multiple terms related to these diseases enriched for the individual candidate PPI networks (section 4.2.1). For example, CRP and GGT1 share enrichment for microvascular angina (182-fold compared to all genes in RHAPSODY) and likewise, we see enrichment in the two candidate PPI networks separately, although this enrichment is only significant for CRP.

In the last section, section 5 –Biomarker potential, our comparisons largely overlap with the results from sections 2 –Experimental evidence and 3 –Disease association, with FAS having the highest scores for antibody count and one of the lowest novelty scores. This indicates that FAS may be a suitable biomarker for T2DM, but has little novelty compared to other candidates. Meanwhile, ENPP7 and IL18R1 have some of the highest novelty scores, but importantly, a relatively high number of antibodies, making these targets suitable as biomarker candidates with high novelty. None of the Slieker *et al* [11] candidates have any known drug

interactions registered in the DrugBank database, while two drugs for CRP are being investigated, both involved in LDL metabolism. Moreover, one and four drugs are approved or sold as nutraceuticals for GGT1 and GPT, respectively.

Comparing the candidates across the five sections, we find some general trends for the nine candidates included here. CRP, a well-known biomarker for T2DM, shows high scores for text mining, disease association scores, and low novelty, as well as interaction with several identified potential biomarkers and high enrichment for gene terms related to inflammation and metabolic disease. Surprisingly, CRP was not found to be significantly associated with T2DM in the RHAPSODY meta-analysis and only a single transcriptomics experiment in pancreas islets shows a significant association between CRP and T2DM. In contrast, CRELD1 and ENPP7 were significantly associated with T2DM in the RHAPSODY meta-analysis, but they are largely unstudied with no association with diabetes according to Open Targets, low text mining scores, and high novelty scores, as well as few or no interaction partners and enriched gene terms. Between these extremes, we find GDF15 to be greatly enriched in the RHAPSODY proteomics analysis, having high association scores for multiple diabetes types and medium-high text mining scores and novelty. GDF15 is, together with its interaction partners, involved in the regulation of food intake and is found to be associated with diabetes and cardiovascular diseases in gene enrichment analysis. Based on these findings, we find CRP to be a poor candidate as a novel biomarker, CRELD1 and ENPP7 to be interesting novel candidates that have largely unknown roles in diabetes, and GDF15 to be a candidate with a strong association with diabetes, but which have not been identified through systematic reviews such as Abbasi *et al* [16] or by biomarker prioritization frameworks such as network expansion.

Comparison to other tools

While multiple online tools exist for the feature analysis of molecular entities in the context of human disease, we have not been able to identify a single biomarker tool that would fit the need for biomarker prioritization for T2DM, though several exist for cancer [30–33] or are aimed more broadly for diseases or drug targets [23,34–37]. Some of the tools require the upload of primary data [37,38] or are made as R packages, requiring a specialized skill set to use [35,36]. We were not able to access all tools that we identified in our search, as they either were behind a paywall [39] or were no longer available online [33,40,41].

These publicly available tools overlap with some of the included features in BALDR, such as protein-protein interaction networks [23,30,33,35] and gene set enrichment analysis [23,30,35]. In addition to these functions, BALDR also provides T2DM-specific experimental data from the RHAPSODY consortium, integration of multiple major databases, and text mining results from 15 million full-text articles. We show the utility of BALDR by comparing six biomarkers for T2DM identified by Slieker *et al* [11] with three highly studied T2DM-biomarkers from Abbasi *et al*[16]. Here, we discuss the molecular insights arising from multiple analyses and highlight candidates according to their novelty as diabetes biomarkers.

Availability and future directions

The BALDR report is provided at https://baldr.cpr.ku.dk/, where the user can request a report on up to 20 biomarker candidates at a time. It includes all published proteomics and transcriptomics data from the RHAPSODY consortium. The report can be downloaded as a .zip-file directly through the website or e-mailed to the user. After compilation, all data processed are deleted immediately. All supplied information, including queried biomarker candidates, is treated as confidential. The website and finished report can be accessed through standard browsers and there are no computational restrictions for running the tool. The report, figures,

and tables can be freely used and adapted by the user. Data from the RHAPSODY consortium can be freely used for academic and industrial purposes. All other data included in the BALDR report is publicly available from the primary sources listed above and should be used according to their individual licenses regarding use and redistribution.

The code for BALDR and its data capture scripts are freely available on GitHub at https://github.com/agnetelundgaard/BALDR. The provided source code for compiling the biomarker matrix that can be adapted to most diseases using EFO labels. However, BALDR has been specifically developed for T2DM, which means that some features, such as full-text text mining and experimental data, are not readily available for other diseases. Other sources for these data types will, therefore, be needed for a complete replication of the report, while protein-protein interaction networks, gene enrichment analysis, and functional information on targets are directly transferable and do not need adaptation to other diseases.

Our hope is that BALDR may serve as a framework for future applications enabling researchers to directly compare molecules of interest using public data without a prerequisite for specialized programming skills.

Supporting information

S1 File. Variables downloaded from the five databases.
(TSV)

S2 File. Known and suggested biomarkers for T2DM.
(TSV)

S3 File. Functional gene set terms from eight databases used in overrepresentation analysis.
(TSV)

S4 File. Example of a BALDR report using nine biomarker candidates from Slieker *et al* [11] and Abbasi *et al.* [16].
(HTML)

Author Contributions

Conceptualization: Agnete T. Lundgaard, Dorte Vistisen, Thomas Sparsø, Giuseppe N. Giordano, Mark Ibberson, Karina Banasik, Søren Brunak.

Data curation: Agnete T. Lundgaard, Frédéric Burdet, Danai Vagiaki, Lisa Cantwell, Timo Röder.

Formal analysis: Agnete T. Lundgaard.

Funding acquisition: Mark Ibberson, Søren Brunak.

Methodology: Agnete T. Lundgaard, Frédéric Burdet, David Westergaard, Timo Röder.

Project administration: Karina Banasik.

Resources: Frédéric Burdet, Mark Ibberson, Søren Brunak.

Software: Agnete T. Lundgaard, Frédéric Burdet, Troels Siggaard, Timo Röder.

Supervision: Dorte Vistisen, Giuseppe N. Giordano, Mark Ibberson, Karina Banasik, Søren Brunak.

Visualization: Agnete T. Lundgaard, Troels Siggaard.

Writing – original draft: Agnete T. Lundgaard, Timo Röder, Dorte Vistisen, Karina Banasik, Søren Brunak.

Writing – review & editing: Agnete T. Lundgaard, Frédéric Burdet, Troels Siggaard, David Westergaard, Danai Vagiaki, Lisa Cantwell, Timo Röder, Dorte Vistisen, Thomas Sparsø, Giuseppe N. Giordano, Mark Ibberson, Karina Banasik, Søren Brunak.

References

1. Obermayer A, Dong L, Hu Q, Golden M, Noble JD, Rodriguez P, et al. DRPPM-EASY: A Web-Based Framework for Integrative Analysis of Multi-Omics Cancer Datasets. Biology (Basel). 2022;11. https://doi.org/10.3390/biology11020260 PMID: 35205126
2. Castellano-Escuder P, Gonzalez-Domnguez R, Carmona-Pontaque F, Andrés-Lacueva C, Sanchez-Pla A. POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis. PLoS Comput Biol. 2021; 17: e1009148. https://doi.org/10.1371/journal.pcbi.1009148 PMID: 34197462
3. Becker AK, Dörr M, Felix SB, Frost F, Grabe HJ, Lerch MM, et al. From heterogeneous healthcare data to disease-specific biomarker networks: A hierarchical Bayesian network approach. PLoS Comput Biol. 2021;17. https://doi.org/10.1371/journal.pcbi.1008735 PMID: 33577591
4. Roglic G, World Health Organization. Global report on diabetes. 2016.
5. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet. 2016; 387: 1513–1530. https://doi.org/10.1016/S0140-6736(16)00618-8 PMID: 27061677
6. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49: D480–D489. https://doi.org/10.1093/nar/gkaa1100 PMID: 33237286
7. Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, et al. TCRD and Pharos 2021: Mining the human proteome for disease biology. Nucleic Acids Res. 2021; 49: D1334–D1346. https://doi.org/10.1093/nar/gkaa993 PMID: 33156327
8. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: A platform for therapeutic target identification and Validation. Nucleic Acids Res. 2017; 45: D985–D994. https://doi.org/10.1093/nar/gkw1055 PMID: 27899665
9. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018; 46: D1074–D1082. https://doi.org/10.1093/nar/gkx1037 PMID: 29126136
10. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010; 28: 1248–1250. https://doi.org/10.1038/nbt1210-1248 PMID: 21139605
11. Slieker RC, Donnelly LA, Lopez-Noriega L, Muniangi-Muhitu H, Akalestou E, Sheikh M, et al. Novel biomarkers for glycaemic deterioration in type 2 diabetes: an IMI RHAPSODY study. medRxiv. 2021; 2021.04.22. https://doi.org/10.1101/2021.04.22.21255625 PMID: 21255625. doi:
12. Solimena M, Schulte AM, Marselli L, Ehehalt F, Richter D, Kleeberg M, et al. Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. Diabetologia. 2018; 61: 641–657. https://doi.org/10.1007/s00125-017-4500-3 PMID: 29185012
13. Wigger L, Barovic M, Brunner AD, Marzetta F, Schöniger E, Mehl F, et al. Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories towards type 2 diabetes. Nature Metabolism 2021 3:7. 2021; 3: 1017–1031. https://doi.org/10.1038/s42255-021-00420-9 PMID: 34183850
14. Sánchez-Archidona AR, Cruciani-Guglielmacci C, Roujeau C, Wigger L, Lallement J, Denom J, et al. Plasma triacylglycerols are biomarkers of β-cell function in mice and humans. Mol Metab. 2021; 54: 101355. https://doi.org/10.1016/J.MOLMET.2021.101355 PMID: 34634522
15. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021; 10: 33. https://doi.org/10.12688/f1000research.29032.2 PMID: 34035898
16. Abbasi A, Sahlqvist AS, Lotta L, Brosnan JM, Vollenweider P, Giabbanelli P, et al. A systematic review of biomarkers and risk of incident type 2 diabetes: An overview of epidemiological, prediction and aetiological research literature. PLoS ONE. Public Library of Science; 2016. p. e0163721. https://doi.org/10.1371/journal.pone.0163721 PMID: 27788146

17. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2016; 14: 61–64. https://doi.org/10.1038/nmeth.4083 PMID: 27892958

18. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016; 536: 41–47. https://doi.org/10.1038/nature18642 PMID: 27398621

19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26: 2190–2191. https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382

20. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57: 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

21. Westergaard D, Stærfeldt HH, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018; 14: 2021. https://doi.org/10.1371/journal.pcbi.1005962 PMID: 29447159

22. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012; 40: D940–D946. https://doi.org/10.1093/nar/gkr972 PMID: 22080554

23. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47: D607–D613. https://doi.org/10.1093/nar/gky1131 PMID: 30476243

24. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33: D433–D437. https://doi.org/10.1093/nar/gki005 PMID: 15608232

25. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019; 47: W199–W205. https://doi.org/10.1093/nar/gkz401 PMID: 31114916

26. R Core Team. R: a Language and Environment for Statistical Computing. R Version 4.0. 1. Vienna, Austria: R Foundation for Statistical Computing; 2020.

27. Baumer B, Udwin D. R Markdown. Wiley Interdisciplinary Reviews: Computational Statistics. Wiley-Blackwell; 2015. pp. 167–177. https://doi.org/10.1002/wics.1348

28. Xie Y. Dynamic Documents with R and knitr. 2nd ed. CRC Press; 2015. Available: https://books.google.dk/books?hl=da&lr=&id=5EQPEAAAQBAJ&oi=fnd&pg=PP1&dq=knitr&ots=IF-gYtBZle&sig=FWNeGdspx3fVs9EkT6NMMvXrLQA&redir_esc=y#v=onepage&q=knitr&f=false

29. Ren K, Russell K. formattable: Create "Formattable" Data Structures. 2021. Available: https://CRAN.R-project.org/package=formattable

30. Ru B, Tong Y, Zhang J. MR4Cancer: A web server prioritizing master regulators for cancer. Bioinformatics. 2019; 35: 636–642. https://doi.org/10.1093/bioinformatics/bty658 PMID: 30052770

31. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaringa R, Rodriguez-Barrientos A, et al. SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis. PLoS One. 2013; 8: 74250. https://doi.org/10.1371/journal.pone.0074250 PMID: 24066126

32. Zheng H, Zhang G, Zhang L, Wang Q, Li H, Han Y, et al. Comprehensive Review of Web Servers and Bioinformatics Tools for Cancer Prognosis Analysis. Front Oncol. 2020; 10: 68. https://doi.org/10.3389/fonc.2020.00068 PMID: 32117725

33. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. Neoplasia. 2007; 9: 166–180. https://doi.org/10.1593/neo.07112 PMID: 17356713

34. Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, et al. PreMedKB: An integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res. 2019; 47: D1090–D1101. https://doi.org/10.1093/nar/gky1042 PMID: 30407536

35. Aguirre-Plans J, Piñero J, Sanz F, Furlong LI, Fernandez-Fuentes N, Oliva B, et al. GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. J Mol Biol. 2019; 431: 2477–2484. https://doi.org/10.1016/j.jmb.2019.02.027 PMID: 30851278

36. Boizard F, Buffin-Meyer B, Aligon J, Teste O, Schanstra JP, Klein J. PRYNT: a tool for prioritization of disease candidates from proteomics data using a combination of shortest-path and random walk algorithms. Sci Rep. 2021; 11: 5764. https://doi.org/10.1038/s41598-021-85135-3 PMID: 33707596

37. Zoppi J, Guillaume JF, Neunlist M, Chaffron S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. BMC Bioinformatics. 2021; 22: 1–14. https://doi.org/10.1186/S12859-020-03921-8/FIGURES/5

38. Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res. 2021 [cited 28 Jun 2021]. https://doi.org/10.1093/nar/gkab382 PMID: 34019663

39. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014; 30: 523–530. https://doi.org/10.1093/bioinformatics/btt703 PMID: 24336805

40. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: Enabling fast and effective prioritization of positional candidates. Bioinformatics. 2006; 22: 773–774. https://doi.org/10.1093/bioinformatics/btk031 PMID: 16423925

41. Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, Deelman E, et al. SPOT: A web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. Nucleic Acids Res. 2010; 38: W201–W209. https://doi.org/10.1093/nar/gkq513 PMID: 20529875