Synteny – a high throughput web tool to streamline causal gene prioritisation and provide insight into protein function

Harry B. Cutler[1,2], Jacqueline Stöckli[1,2], Søren Madsen[1,2], Stewart W.C. Masson[1,2], Oliver K. Fuller[1,2], Tom Buckingham Shum[2], David E. James[1,2,3,4]

[1] School of Life and Environmental Sciences, University of Sydney, Camperdown, New South Wales, Australia.
[2] Charles Perkins Centre, University of Sydney, Camperdown, New South Wales, Australia.
[3] School of Medical Sciences, University of Sydney, Camperdown, New South Wales, Australia.
[4] Correspondence: david.james@sydney.edu.au

Summary

Disease-associated genes can now be identified more readily than their functions can be understood. However, determining which candidates to take forward for functional validation remains a substantial challenge. Assessing whether candidate genes are associated with human phenotypes is a promising approach, although a systematic method to integrate the results of studies in model systems with human genetic data does not currently exist. Here we present *Synteny*, a powerful tool that enhances the accessibility of human genetic data and enables interactive visualisation of the associations between genes and phenotypes (https://bigproteomics.shinyapps.io/Synteny/). *Synteny* assists experimental biologists prioritise genes for further investigation, which we demonstrate by interrogating a locus from a published genetic analysis of lifespan in Diversity Outbred mice. To illustrate how *Synteny* provides unique insights into protein function, we applied it to analyse genes within the solute carrier (SLC) superfamily. Despite many members of the SLC superfamily remaining uncharacterised, *Synteny's* unbiased genetic analysis identified two potentially novel urate transporters, showcasing *Synteny* as a powerful tool for identifying functionally important proteins.

Keywords

Disease-associated genes, candidate prioritisation, causal genes, human genetic data, proteomics, transcriptomics, systems biology

Introduction

Advances in the availability and performance of omics-based techniques such as proteomics and transcriptomics have revolutionised the study and understanding of human disease. An important task now is to develop computational tools to assist both experts and non-experts to extract biological meaning from such data. The complexity of this task cannot be understated.

The first challenge lies in the scale of the data. Omics studies frequently identify hundreds of candidate genes with strong molecular links to phenotype, forcing investigators to decide which to pursue experimentally. Although searching the literature can certainly assist in this process, and several tools for improving the efficiency of literature searching are available[1–3], many candidates may not have been studied before and so these tools will only provide insight for a subset of genes[4–6]. Even in rare cases where all candidates are functionally annotated, proteins can exhibit distinct functions depending on the context in which they are being studied and so their relevance to a new phenotype may not be immediately clear. Faced with the challenge of generating hypotheses and designing experiments to validate the functions of uncharacterised proteins, investigators battle an inherent bias towards

selecting candidates with existing links to their phenotype of interest [6,7]. This 'streetlight' bias – searching for answers where the light is brightest rather than venturing into the relative unknown – impedes the rate at which new breakthroughs can be made and signals an unmet need for more robust tools for unbiased candidate prioritisation.

The second obstacle lies in interpreting the complexity within the data. Many omics-based studies identify genes through differential expression or correlation analyses. However, these approaches are typically unable to ascertain whether the identified genes are causal or consequential. The development of systematic approaches to identify causal genes is therefore critical for increasing the success rate of future validation studies as most differentially expressed genes appear to be downstream of phenotypic changes and are not causal[8].

One approach that is becoming popular is to integrate omics data with findings from human genetic studies. Genetic variants are causal by definition and so the confidence in a gene's causality increases when genetic signals influencing phenotypic variation and gene expression colocalise. Genetic and phenotypic data are now available across large human populations and several studies using model systems have demonstrated success in applying such an integrated approach. For example: Keller *et al* leveraged human genetic data to demonstrate that loci detected in a screen of insulin secretion in Diversity Outbred mice were enriched for diabetes-related traits in humans[9]; Li *et al* integrated human genetic data and mouse gene co-expression networks to identify *SESN1* as a regulator of hepatic cholesterol synthesis[10]; and Parker *et al* identified *PSMD9* as a regulator of hepatic lipotoxicity in both mice and humans[11]. These examples highlight the potential benefits of developing systematic approaches to interrogating human genetic data and suggest that harmonising the results of studies in humans and model systems may be essential to maximising biological insight. However, these approaches require specialist computational skills and may be challenging for many researchers to integrate with their own data.

To provide all researchers the ability to assess their findings in the context of human genetic data, we developed *Synteny* (https://bigproteomics.shinyapps.io/Synteny/). *Synteny* is a powerful interface to the Association to Function (A2F) database (https://a2f.hugeamp.org/) that, as of February 2025, reports summary statistics for 898 genetic studies and 1175 phenotypes. *Synteny* was inspired by an unmet need to not only identify syntenic regions between species, but syntenic function too.

Results

Synteny powers improved access to human genetic evidence

As few as 14% of studies consider human genetic data when selecting genes for mechanistic study[12]. While lack of expertise in interpreting the results of human genetic studies certainly provided a barrier to data integration in the past, more recent approaches such as the Human Genetic Evidence (HuGE) framework have opened the doors to those unfamiliar with the fine details of genetic mapping[12]. HuGE scores provide a single measure for how strongly a given gene associates with a phenotype. They are calculated by considering the effects of both rare and common genetic variation, the proximity of a gene to the peak genetic association in genome wide association studies and the impact of coding mutations detected in whole exome sequencing. HuGE scores aggregate genetic data across many independent studies, providing a powerful and unbiased means by which to assess how a gene of interest might relate to human physiology across many phenotypes. For example, consistent with the known role of *PPARG* as a master regulator of adipogenesis, inspecting the HuGE scores for *PPARG* quickly reveals very strong (HuGE ≥ 30) links to body fat percentage and muscle fat infiltration (Fig 1A). Accessing HuGE scores for individual genes is now straightforward thanks to the efforts of those behind the

Association to Function (A2F) and Type 2 Diabetes Knowledge Portals[13]. However, modern omics studies may identify hundreds of genes, and the current restriction of these portals to single-gene queries limits the ability of these portals to be interrogated in the context of high throughput hypothesis-generating studies (Fig 1B). Many candidates from such studies have low HuGE scores and so fundamental benefits exist in designing an initial screening tool to accelerate the prioritisation of genes for further study.

We developed *Synteny* to address this challenge. *Synteny* greatly enhances the accessibility of human genetic evidence by facilitating bulk querying of the A2F database and assisting researchers to identify high-priority candidates for mechanistic follow-up (Fig 1C). *Synteny* is compatible with data all common model species; gene names from these species are automatically mapped to their human orthologs using the Alliance for Genome Resources database[14]. All results generated in *Synteny* can be exported for analysis in external software, and result visualisation plots can be downloaded as high-resolution vector PDFs.

The primary function of *Synteny* is to enable users to acquire human genetic evidence for many genes simultaneously. Interactive visualisation of the results enables rapid identification of genes with the strongest associations to human phenotypes. To demonstrate this, we employed *Synteny* to analyse a list of 52 genes that Dornbos et al originally used to assess human genetic evidence for genes hypothesised to relate to glucose metabolism[12] (Fig 1D). While performing this analysis without *Synteny* would likely have taken upwards of an hour by manually querying each gene in the A2F database, it took only 1 min and 20 seconds using *Synteny*, and it is immediately clear which genes possess the strongest links to human disease. Investigation of HuGE scores in this way frequently reveals well-studied associations, however, the true value of this analysis lies in highlighting less well appreciated associations that might have gone unnoticed otherwise. For example, while the link between *PTEN* and systemic glucose homeostasis is well-established[15,16], we were surprised by the link between *CPT1A* and platelet number. CPT1A plays a central role in transporting fatty acids into the mitochondria for oxidation[17,18] and, consistent with human genetic evidence, recent studies have revealed that the contribution of fatty acids to ATP production in platelets is likely greater than currently appreciated[19,20]. Whether investigation of this pathway might provide novel insight into platelet pathology is unclear, however it emphasises how increasing access to human genetic data can lead to enlightened hypotheses.

The second function of *Synteny* is to provide insight into protein function. Although many of the genes in Figure 1 are well studied, this is unlikely to be the case for many genes identified in many omics studies. One approach is to adopt a guilt-by-association style strategy and hypothesise that genes with similar functions should be associated with similar human phenotypes. Such strategies are implemented in several established tools including WGCNA[21], GeneBridge[22] and QENIE[23], all of which assign putative functions to genes by means of correlations or co-expression networks. To demonstrate the utility of this approach, we correlated HuGE scores across the 52 genes from Fig 1D and identified several gene pairs with shared phenotypic associations (Fig 1E). For example, *FFAR4* and *PIK3CA* both associated strongly with red blood cell count. Interestingly, both FFAR4 and PI3KCA can influence the levels of phosphoinositide (PI) signalling lipids[24,25] and developing red cells require high rates of incorporation of PI lipid substrates[26]. Thus, the correlation between the HuGE scores of these two genes might stem from their shared ability to regulate PI levels. When applied to functionally uncharacterised genes, therefore, this approach provides unbiased insight into protein function.

Not all findings from studies in model systems should be expected to translate to humans and so understanding the extent to which these systems recapitulate the mechanisms of human disease is critical to enabling the most relevant model systems to be prioritised in future. For example, the incorporation of genetic diversity is a promising approach by which to attain new biological insight[27,28]. Human Phenotype Enrichment (HPE) is only possible in

Synteny and enables unbiased assessment of how strongly findings from studies in model systems align with human genetic links between genes and phenotypes. Specifically, *Synteny's* HPE compares HuGE scores for a user-supplied gene set against those resulting from random sampling of the genome (or a user-supplied background gene set). This enables the gene set's collective strength of association to be evaluated since functionally enriched gene sets are expected to have higher HuGE scores. Consistent with the interpretation of Dornbos *et al*[12] , HPE of the 52 genes in Fig 1D failed to detect significant enrichment for glycaemic traits or those relating to diabetic complications that would be expected if these genes were supported by human genetic evidence (Fig 1F). Instead, these genes were enriched for early life Body Mass Index (BMI), Alzheimer's disease (AD) and age-related macular degeneration (AMD). As has been argued by Dornbos[12] and others previously, this emphasises the importance of interrogating results in the context of human genetics to maximise the extent to which advances in basic research can be translated into human therapeutic breakthroughs.

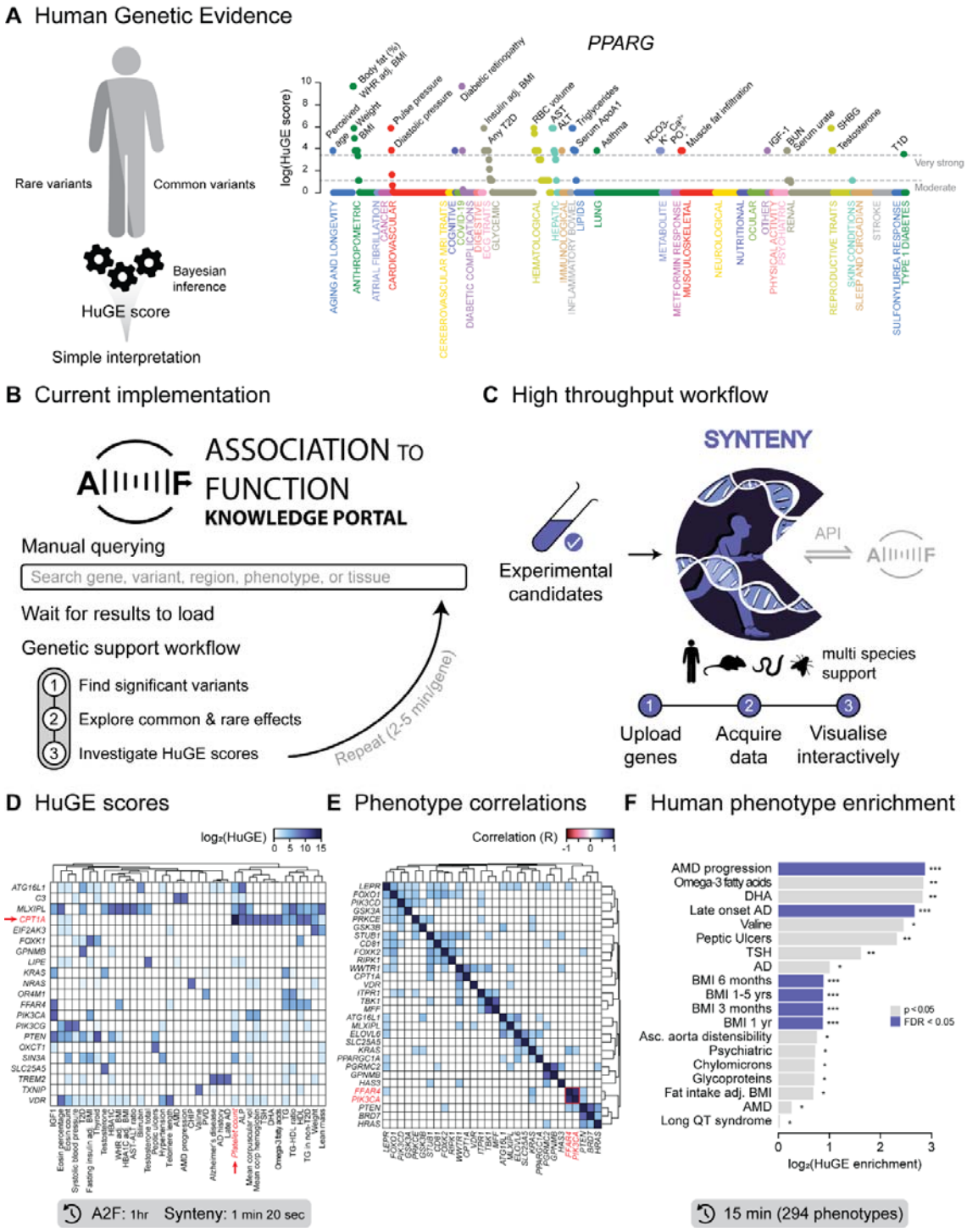


Figure 1 – Synteny powers improved access to human genetic evidence. (A) Human genetic evidence (HuGE) scores aggregate genetic data from many studies to summarise the strength of association between genes and phenotypes[12], enabling gene function to be rapidly assessed. Example of HuGE score plot linking *PPARG* to 1175 phenotypes. **(B)** Schematic of the current implementation of the HuGE score framework via the Association to Function web portal. **(C)** Schematic illustrating the high throughput implementation in *Synteny*. API = application programming interface. **(D-F)** Examples of data visualisation and analysis in *Synteny* for 52 genes hypothesised to be linked to glucose metabolism[12]. **(D)**

Heatmap showing associations between genes and human phenotypes. **(E)** Correlation heatmap of HuGE scores. **(F)** Results of human phenotype enrichment. Asterisks represent unadjusted p values.

Synteny streamlines causal gene prioritisation

Determining which candidate genes are most likely to be causal is a significant hurdle in many studies. Although many have suggested that this process could be streamlined by cross-referencing results with human genetic datasets, few scalable approaches have been reported. *Synteny* creates a systematic web-based pipeline for causal gene prioritisation. We demonstrate this by applying *Synteny* to a recent study of longevity in Diversity Outbred (DO) mice where Di Francesco *et al*[29] reported a genome-wide significant QTL for lifespan mapping to chromosome 18 at 21.55 Mb, but did not pursue a causal gene (Fig 2A).

We began our analysis at this point, seeking to leverage human genetic evidence to provide insight into which candidate was the most likely to be causal. The chromosome 18 locus identified by Di Francesco *et al* contained 23 genes. We interrogated these in *Synteny* across 1175 phenotypes, identifying 52 phenotypes with at least moderate genetic support (HuGE score ≥ 3) and 19 phenotypes with very strong support (HuGE ≥ 30). Among these very strongly supported phenotypes were BMI, Type 2 Diabetes and, strikingly, telomere length (TL, Fig 2B). Telomere length is one of the hallmarks of longevity[30] and is genetically correlated with parental longevity in the A2F database ($R = 0.28$, $p < 1.46 \times 10^{-8}$), consistent with the hypothesis that one or more of the genes in the chromosome 18 locus regulate lifespan in humans. Of the 23 genes in the locus, only 2 possessed very strong genetic evidence linking them to telomere length: *TRAPPC8* and *FHOD3*. *TRAPPC8* was most strongly linked to HDL cholesterol, TL and plasma triglyceride levels, whereas *FHOD3* demonstrated compelling genetic evidence (HuGE ≥ 350) linking it to bone mineral density, heart rate, the diabetes biomarker HBA1C and an array of cardiac phenotypes (left and right ventricular ejection fractions and dilated cardiomyopathy) in addition to TL. Interestingly, dilated cardiomyopathy exhibits a strong genetic correlation to parental lifespan ($R = -0.43$, $p < 2.4 \times 10^{-10}$), suggesting that *FHOD3* might regulate longevity by modifying cardiac function. Thus, use of *Synteny* narrowed the selection from 23 genes to 2 and provided potential mechanistic insight for *FHOD3*, enabling further efforts to be devoted to comprehensive review of the protein function literature for *FHOD3* and *TRAPPC8*.

TRAPPC8 was first identified as the most proximal component of a TRAPP complex that is required for normal ATG9 trafficking and the induction of autophagy[31]. Since then, it has also been found to interact with TMEM131, deficiency of which activates the endoplasmic reticulum (ER) stress response[32]. Both autophagy and the ER stress response are well-established regulators of longevity[33–36], providing a plausible link between genetic variation in *TRAPPC8* and human lifespan. Consistent with links to heart disease in human genetic data, a muscle-specific isoform of FHOD3 regulates actin filament polymerisation in cardiac muscle[37]. In mice, homozygous deletion of *FHOD3* is embryonic lethal, heterozygous deletion results in juvenile lethality and adult-induced deletion resulted in cardiac dilation and dysfunction 50 weeks after deletion induction[38], providing a firm mechanistic link between FHOD3, cardiac function and lifespan. Given that both *TRAPPC8* and *FHOD3* possessed plausible mechanistic links to lifespan, we turned to publicly available data from Diversity Outbred mice to determine if one of the two genes could be further prioritised.

Assuming that the effect of the chromosome 18 QTL is mediated by a change in gene expression, we sought to identify which of *TRAPPC8* or *FHOD3* most closely aligned with the founder allele contributions to lifespan. Importantly, it is possible to estimate the contribution of each founder to a genetic signal as DO mice originated from the outbreeding of 8 inbred founder strains and all of these have been fully sequenced[39,40]. We obtained publicly available gene expression QTL (eQTL) founder contribution data for *FHOD3* and

TRAPPC8 using the QTLViewer[41], and then correlated the founder effects of gene expression in each tissue with those of longevity (Fig 2C). This identified *FHOD3* as being more closely associated with longevity than *TRAPPC8*, with founder effects significantly correlating with lifespan in the hippocampus (Fig 2D). We performed this analysis across all available tissues since the regulation of gene expression can be highly tissue specific and it is not clear which tissue is most relevant for lifespan. However, we cannot rule out the possibility that *TRAPPC8* would be more strongly implicated in tissues not currently included in the QTLViewer database.

This example demonstrates a systematic pipeline with which to prioritise candidates for validation experiments. The use of *Synteny* as a filtering step was critical here as it would not have been possible to comprehensively review the protein function literature for all 23 genes. It also highlights the power of maintaining an unbiased approach to candidate selection since, if we had only looked for human genetic evidence relating to aging and longevity-related phenotypes, we would have missed important information relating to potential mechanisms. Obtaining these insights is pivotal for designing the most informative validation experiments.

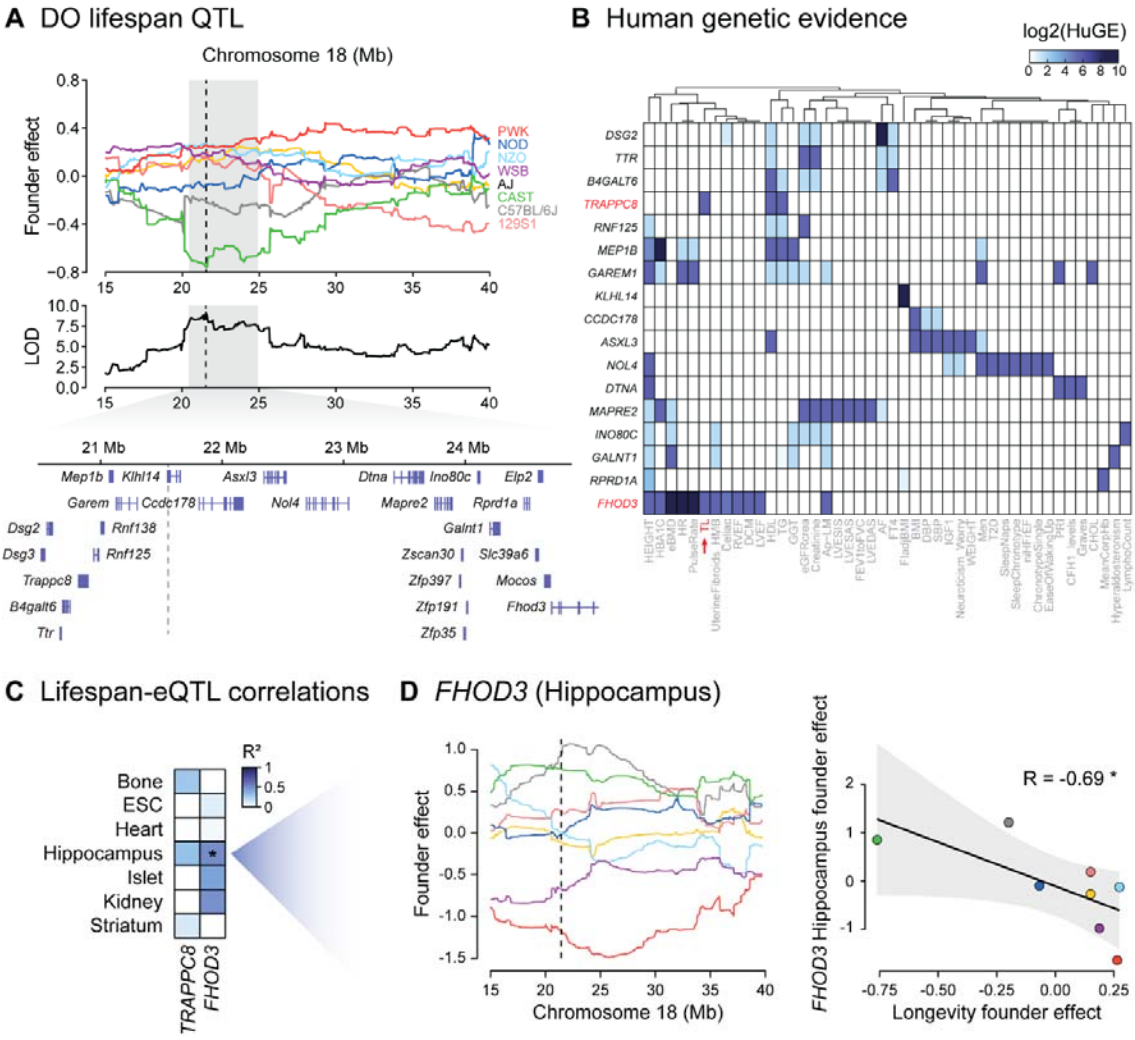


Figure 2 – Synteny streamlines causal gene prioritisation. (A) Quantitative trait locus for lifespan detected by Di Francesco et al in a study of Diversity Outbred mice[29]. Founder contributions (top panel), longevity LOD score (middle panel) are shown between 15-40 Mb

on chromosome 18. Genes within 95% confidence interval are shown (bottom panel). **(B)** Heatmap of HuGE scores for genes contained within the chromosome 18 QTL that have at least one very strong HuGE score (HuGE ≥ 30). Genes are ordered along the y-axis by their position on the chromosome. **(C)** Correlations between founder effects for the lifespan QTL and those for *FHOD3* and *TRAPPC8* eQTLs in publicly available data contained within the QTLViewer[41]. ESC = embryonic stem cells. **(D)** Specific correlation between the *FHOD3* eQTL in kidney and the founder effects for the lifespan QTL.

Synteny provides insight into protein function

Less than half of human genes have been studied mechanistically[7,22,42]. To evaluate whether human genetic data can be used to provide insight into protein function, we used *Synteny* to investigate genes within the solute carrier (SLC) superfamily, containing over 400 genes. We chose this family since, despite some members playing crucial roles in many aspects of biology, much remains to be learnt about the substrates and biological functions of many of these proteins[43].

Human phenotype enrichment (HPE) offers an unbiased approach to determine how a list of genes contribute to human physiology. Applying this analysis to the SLC proteins revealed that SLC proteins are enriched for roles in modulating 24 human traits (Fig 3A). Consistent with the predominant functions of SLCs as transporters, the phenotypes featured circulating levels of a variety of metabolites such as creatinine, urate and sodium, and even extended to overt disease conditions such as ‘chronic kidney disease’ and ‘epilepsy with sclerosis’.

To identify putative new functions for SLC members, we identified genes with similar phenotypic associations by correlating HuGE scores between genes. Using a very stringent threshold for correlations with $R > 0.95$ and $p < 0.01$, we identified 6 clusters that contained more than 2 SLC proteins (Fig 3B). Each SLC cluster was enriched for at least one human trait. Cluster 4 was enriched for genes associating with serum urate levels and, strikingly, 3 of the 5 genes in this cluster (*SLC2A9*, *SLC22A12*, *SLC22A13*) have previously been shown to actively transport urate[44–46]. This raises a question as to whether the other 2 SLC members in this cluster, *SLC5A2* and *SLC5A9*, also function as urate transporters.

SLC5A2, also known as SGLT2, is a sodium dependent glucose transporter that is highly expressed in the kidney and is a well-established target in the treatment of type 2 diabetes[47,48]. Inhibition of SGLT2 prevents glucose retrieval in the kidney and leads to disposal of glucose in the urine. However, direct evidence linking it to serum urate levels also exists. A coding variant in *SLC5A2* (c.469-1G>A) has been linked to uricosuria[49], a condition with high urate levels in the urine. Additionally, pharmacological inhibition of SGLT2 decreases the re-absorption of urate from the urine[50]. It has been suggested that SGLT2 inhibitors induce uricosuria via an indirect mechanism whereby elevated urinary glucose activates enhanced urate transport by GLUT9 (*SLC2A9*)[49]. However, GLUT9 has previously been shown to be dispensable for the effects of SGLT2 inhibition on hyperuricosuria[50]. Although altered glucose concentrations have also been proposed to modulate the activity of *SLC22A12* (URAT1) [49,51], the predominant urate transporter in the renal tubule, we were unable to find any studies explicitly testing the effects of increasing glucose concentration on the activity of URAT1. Collectively, these data indicate a potential direct role for SGLT2 in urate transport, although future studies are required to test this hypothesis.

SLC5A9, also known as SGLT4, has previously been shown to function as a sodium-dependent mannose and fructose transporter in the small intestine where it contributes to the absorption of sugars from the diet[52,53], although little else is known about its function. Combined with the fact that approximately 30% of uric acid is excreted into the small intestine, the known overlap between urate and sugar transport[44] suggests that *SLC5A9* might also transport urate. Given that enhancing intestinal urate disposal has been proposed

as a therapeutic target for chronic kidney disease[54], it will be important for future studies to clarify the role of *SLC5A9* in dictating serum urate levels in humans.

These examples highlight how improving access to human genetic data can provide insight into protein function and emphasise the utility of *Synteny* in providing information that is essential to the design of follow-up experiments.

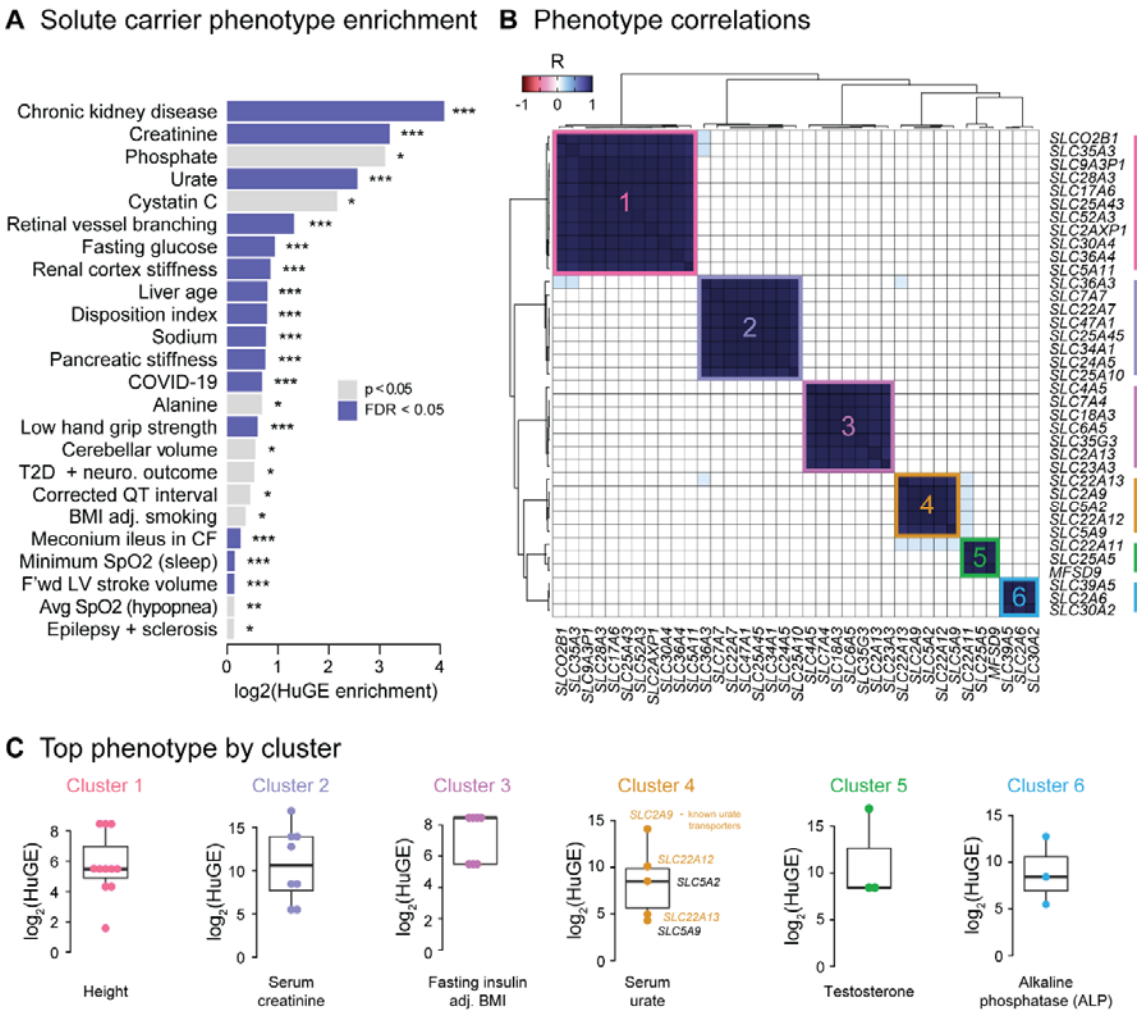


Figure 3 – Synteny provides insight into protein function. (A) Results of human phenotype enrichment for 413 members of the solute carrier (SLC) superfamily, only possible using Synteny. Asterisks represent unadjusted p values. **(B)** Correlation heatmap of HuGE scores for SLC genes with more than one R > 0.95 p < 0.01. Clusters contain SLC genes with similar phenotypic associations. **(C)** Top scoring phenotype for each SLC cluster.

Discussion

The development of improved methods to integrate human genetic data into preclinical studies is essential to maximising the efficiency with which basic research can be translated into improved therapies. Combined with existing bioinformatic approaches to identify genes of interest, *Synteny* forms a systematic pipeline to assist researchers to identify candidates with the greatest molecular support for their involvement in human disease.

Synteny builds upon an existing framework for communicating human genetic evidence. HuGE scores strike an important balance between accessibility for the general community and reporting human genetic associations with absolute fidelity to all available data. Accessed via *Synteny*, the HuGE score framework facilitates rapid candidate screening and causal inference. It is important to recognise that the A2F and Type 2 Diabetes Knowledge Portals contain more data than can reasonably be included within *Synteny*. While *Synteny* greatly simplifies the process of acquiring HuGE scores and determining the genetic correlations between human phenotypes, it was not designed to facilitate detailed inspection of specific candidate genes. To address this limitation, *Synteny* automatically generates hyperlinks to the A2F database for all genes and phenotypes of interest to encourage users to continue to perform in-depth analyses using existing portals once strong candidates have been identified.

Although we have emphasised the utility of *Synteny* in assisting researchers working in model systems, it also works in reverse in the sense that it may help to prioritise human genes for further study. Specifically, human genome-wide association studies (GWAS) have been very successful in identifying candidate loci, however, pinpointing the causal gene/s within these loci remains a persistent challenge. This raises a question as to whether the mechanisms underlying these signals may be better understood by conducting future studies in model systems[55]. Many single nucleotide polymorphisms occur in non-coding regions and, despite being assumed to affect transcription, only 8% of disease quantitative trait loci have been linked to altered gene expression in humans[56]. Although cell-type specificity and/or gene-environment interactions are likely critical in determining the contexts in which genetic variation influences gene expression, these are difficult to study in humans. In contrast, studies in model systems benefit from strict environmental control, access to all cell types and deep phenotyping. The question, therefore, is not whether it is still relevant to study model systems, but how best to integrate them with human studies so that we can better understand fundamental aspects of biology like gene-environment interactions.

Synteny facilitates several unique analyses that are not possible elsewhere. Human phenotype enrichment (HPE) is an unbiased method for assessing the functions of gene sets in humans. Here we demonstrated the utility of HPE to firstly assess how closely candidate genes for glucose metabolism are supported by human genetic evidence (Fig 1F) and, later, to identify overarching trends in the functions of proteins in the solute carrier superfamily (Fig 3A). Many other uses are possible. For example, investigators could perform HPE on differentially expressed genes to determine whether human genetic evidence supports a causal role for the identified proteins. This would determine whether the differentially expressed gene set was more associated with human phenotypes than expected by chance. Alternatively, biomarkers for rare conditions could be searched to identify other associated phenotypes that may have otherwise gone undetected. This has important implications for the study of rare diseases where knowledge of these secondary phenotypes may provide means by which to improve diagnosis or treatment and further emphasises the importance of using *Synteny* to conduct unbiased analyses across all available phenotypes.

Synteny also provides functional insights for uncharacterised proteins. By searching for solute carriers with consistent associations to human traits, we identified *SLC5A2* (SGLT2)

and *SLC5A9* as putative urate transporters. Urate transport in the kidney and small intestine are complicated by the interconnected mechanisms of many interfering transporters. For example, although urate excretion is significantly elevated by SGLT2 inhibition, it remains uncertain whether this occurs by a direct effect on the transport of urate by SGLT2 or an indirect effect of altered urinary glucose levels on the activity of other urate transporters. Adequately testing this would require detailed kinetic experiments to determine the K_m of SGLT2 for urate compared to other substrates as well as experiments to assess the inhibitory capacity of increasing glucose concentrations on urate transport by URAT1 or SLC2A9. Such experiments would need to be performed in cell lines with correct polarisation and expression of several transporters and are therefore outside the scope of this work. Additionally, regardless of whether *SLC5A2* and *SLC5A9* in fact transport urate themselves, genetic or pharmacological inhibition, at least in the case of *SLC5A2*, is sufficient to modulate serum urate levels and this provides effective demonstration of the use of human genetic evidence to group genes by their physiological functions.

These results provide a strong proof of concept for the use of human genetic data in establishing a systematic approach towards prioritising candidate genes and gaining insight into protein function. *Synteny* greatly simplifies the integration of data from humans with that from other systems and provides a user-friendly interface to maximise the biological insight that can be drawn from human genetic datasets (https://bigproteomics.shinyapps.io/Synteny/).

Resource availability

Synteny is freely available at the following address: https://bigproteomics.shinyapps.io/Synteny/. Code is open-source and available in the supplementary materials. This can be run locally using RStudio.

Acknowledgements

This work was supported by an Australian Research Council Laureate Fellowship (to DEJ) and an Australian Government Research Training Program Scholarship (to HBC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the ARC.

We would like to thank several colleagues for their constructive feedback: Marcus Seldin for suggestions relating to enabling broader use of the platform by incorporating the Alliance for Genome Resources database; Alice Williamson for suggestions relating to heatmap visualisation settings and improving the documentation for HuGE correlation analysis; and Rob Williams for suggestions relating to the inclusion of metadata in exported results. These suggestions greatly improved the user experience of *Synteny,* and we welcome future suggestions from the community.

Author contributions

Conceptualisation, HBC, JS, SM, SWCM, OKF, DEJ; Methodology, HBC; Formal Analysis, HBC; Data curation, HBC; Writing – Original Draft, HBC; Writing – Review & Editing, HBC, JS, SM, SWCM, OKF, DEJ; Visualisation, HBC, TBS; Supervision, DEJ; Funding Acquisition, DEJ.

Declaration of interests

The authors declare no competing interests.

Supplemental information

Supplementary file 1: zip file containing source code required to run *Synteny* locally.

Figure legends

Figure 1 – Synteny powers improved access to human genetic evidence. (A) Human genetic evidence (HuGE) scores aggregate genetic data from many studies to summarise the strength of association between genes and phenotypes[12], enabling gene function to be rapidly assessed. Example of HuGE score plot linking *PPARG* to 1175 phenotypes. (B) Schematic of the current implementation of the HuGE score framework via the Association to Function web portal. (C) Schematic illustrating the high throughput implementation in *Synteny*. API = application programming interface. (D-F) Examples of data visualisation and analysis in *Synteny* for 52 genes hypothesised to be linked to glucose metabolism[12]. (D) Heatmap showing associations between genes and human phenotypes. (E) Correlation heatmap of HuGE scores. (F) Results of human phenotype enrichment. Asterisks represent unadjusted p values.

Figure 2 – Synteny streamlines causal gene prioritisation. (A) Quantitative trait locus for lifespan detected by Di Francesco et al in a study of Diversity Outbred mice[29]. Founder contributions (top panel), longevity LOD score (middle panel) are shown between 15-40 Mb on chromosome 18. Genes within 95% confidence interval are shown (bottom panel). (B)

Heatmap of HuGE scores for genes contained within the chromosome 18 QTL that have at least one very strong HuGE score (HuGE ≥ 30). Genes are ordered along the y-axis by their position on the chromosome. (C) Correlations between founder effects for the lifespan QTL and those for *FHOD3* and *TRAPPC8* eQTLs in publicly available data contained within the QTLViewer[41]. ESC = embryonic stem cells. (D) Specific correlation between the *FHOD3* eQTL in kidney and the founder effects for the lifespan QTL.

Figure 3 – Synteny provides insight into protein function. (A) Results of human phenotype enrichment for 413 members of the solute carrier (SLC) superfamily, only possible using Synteny. Asterisks represent unadjusted p values. (B) Correlation heatmap of HuGE scores for SLC genes with more than one $R > 0.95$ $p < 0.01$. Clusters contain SLC genes with similar phenotypic associations. (C) Top scoring phenotype for each SLC cluster.

Methods

Statistical analyses

All analysis and data visualisation were conducted using the R programming environment[57]. Unless otherwise stated, correlations were assessed by Pearson correlation analysis. Significance is represented, with a p value < 0.05 by *, < 0.01 by **, < 0.001 by *** and < 0.0001 by ****.

Human phenotype enrichment

Human genetic evidence (HuGE) scores were downloaded for each gene. To assess enrichment for a particular phenotype, the same number of genes were randomly selected from the Association to Function database (for naïve enrichment) or randomly from a set of background genes (for background-corrected enrichment). This was repeated for 10,000 permutations to enable the average HuGE score in the user-defined gene set to be compared to the sampled distribution/s. An enrichment score was calculated as the mean fold change of the average scores, and significance was determined by calculating the proportion of sampled scores exceeding that of the user-defined set. Multiple corrections were performed using a false-discovery rate (FDR) correction to account for the number of phenotype tests.

Synteny

Web server implementation

Synteny is implemented as a Shiny application using the R programming language and is hosted on a shinyapps.io server.

Privacy and security

Gene lists are only uploaded if the user accesses *Synteny* via the web server. No user data are retained following session termination. Secure HTTPS connections are used to transfer data to and from the server. The country in which each user is located is logged using a google analytics cookie. This data is used solely to monitor global adoption of the tool and ensure server requirements are sufficient to meet demand. No other data is recorded.

Local installation

The code required to run *Synteny* locally is available in the supplementary materials. After installing the required R packages, *Synteny* will still require an internet connection to connect to the Association to Function application programming interface. No data will be uploaded to the *Synteny* shinyapps.io server if running locally.

Packages and databases

Synteny maps gene names from any organism to human orthologs using the Alliance for Genome Resources database[14]. Human genetic evidence (HuGE) scores are retrieved via an application programming interface connected to the Association to Function database using the jsonlite package[58]. Static heatmaps are created using the ComplexHeatmap package[59] and interactive plots are created using a combination of ggplot2[60] and plotly[61].

Exporting results

Results and figures can be explored and visualised within *Synteny* itself or exported for analysis in the users' software of choice. Data is exportable as excel sheets (.xlsx) and figures are exportable as full resolution vector PDFs.

References

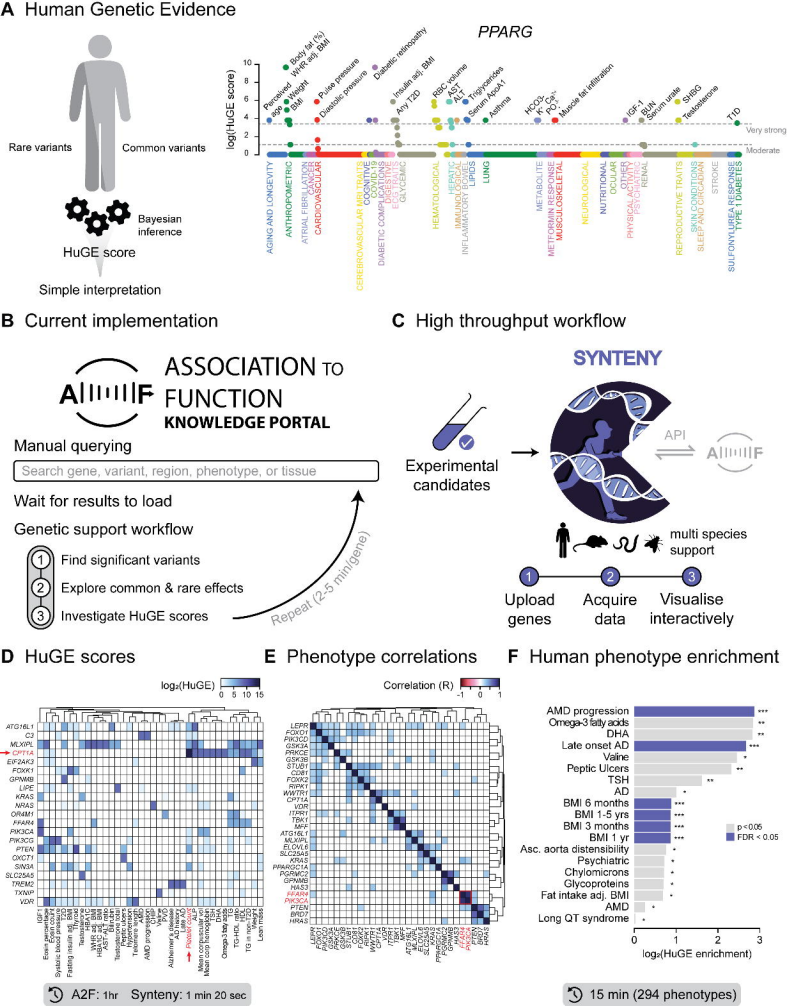
1. Gunturkun, M.H., Flashner, E., Wang, T., Mulligan, M.K., Williams, R.W., Prins, P., and Chen, H. (2022). GeneCup: mining PubMed and GWAS catalog for gene-keyword relationships. G3 (Bethesda) *12*. https://doi.org/10.1093/G3JOURNAL/JKAC059.
2. Soto, A.J., Przybyła, P., and Ananiadou, S. (2019). Thalia: semantic search engine for biomedical abstracts. Bioinformatics *35*, 1799–1801. https://doi.org/10.1093/BIOINFORMATICS/BTY871.
3. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., and Leser, U. (2012). GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Res *40*, W585. https://doi.org/10.1093/NAR/GKS563.
4. Müller, J.B., Geyer, P.E., Colaço, A.R., Treit, P. V., Strauss, M.T., Oroshi, M., Doll, S., Virreira Winter, S., Bader, J.M., Köhler, N., et al. (2020). The proteome landscape of the kingdoms of life. Nature 2020 582:7813 *582*, 592–596. https://doi.org/10.1038/s41586-020-2402-x.
5. Hutchison, C.A., Chuang, R.Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., et al. (2016). Design and synthesis of a minimal bacterial genome. Science (1979) *351*. https://doi.org/10.1126/SCIENCE.AAD6253/SUPPL_FILE/AAD6253-HUTCHISON-SM.PDF.
6. Kustatscher, G., Collins, T., Gingras, A.C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K.S., Lundberg, E., Marcotte, E.M., Ralser, M., et al. (2022). Understudied proteins: opportunities and challenges for functional proteomics. Nat Methods *19*, 774–779. https://doi.org/10.1038/S41592-022-01454-X.
7. Edwards, A.M., Isserlin, R., Bader, G.D., Frye, S. V., Willson, T.M., and Yu, F.H. (2011). Too many roads not taken. Nature 2011 470:7333 *470*, 163–165. https://doi.org/10.1038/470163a.
8. Porcu, E., Sadler, M.C., Lepik, K., Auwerx, C., Wood, A.R., Weihs, A., Sleiman, M.S.B., Ribeiro, D.M., Bandinelli, S., Tanaka, T., et al. (2021). Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nature Communications 2021 12:1 *12*, 1–9. https://doi.org/10.1038/s41467-021-25805-y.
9. Keller, M.P., Rabaglia, M.E., Schueler, K.L., Stapleton, D.S., Gatti, D.M., Vincent, M., Mitok, K.A., Wang, Z., Ishimura, T., Simonett, S.P., et al. (2019). Gene loci associated with insulin secretion in islets from nondiabetic mice. J Clin Invest *129*, 4419–4432. https://doi.org/10.1172/JCI129143.
10. Li, Z., Votava, J.A., Zajac, G.J.M., Lamming, D.W., Liang, C.-, Yen, E., and Parks Correspondence, B.W. (2020). Integrating Mouse and Human Genetic Data to Move beyond GWAS and Identify Causal Genes in Cholesterol Metabolism. Cell Metab *31*, 741-754.e5. https://doi.org/10.1016/j.cmet.2020.02.015.
11. Parker, B.L., Calkin, A.C., Seldin, M.M., Keating, M.F., Tarling, E.J., Yang, P., Moody, S.C., Liu, Y., Zerenturk, E.J., Needham, E.J., et al. (2019). An integrative systems genetic analysis of mammalian lipid metabolism. Nature *567*, 187. https://doi.org/10.1038/S41586-019-0984-Y.
12. Dornbos, P., Singh, P., Jang, D.K., Mahajan, A., Biddinger, S.B., Rotter, J.I., McCarthy, M.I., and Flannick, J. (2022). Evaluating human genetic support for hypothesized metabolic disease genes. Cell Metab *34*, 661–666. https://doi.org/10.1016/j.cmet.2022.03.011.
13. Costanzo, M.C., von Grotthuss, M., Massung, J., Jang, D., Caulkins, L., Koesterer, R., Gilbert, C., Welch, R.P., Kudtarkar, P., Hoang, Q., et al. (2023). The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits. Cell Metab *35*, 695-710.e6. https://doi.org/10.1016/J.CMET.2023.03.001.

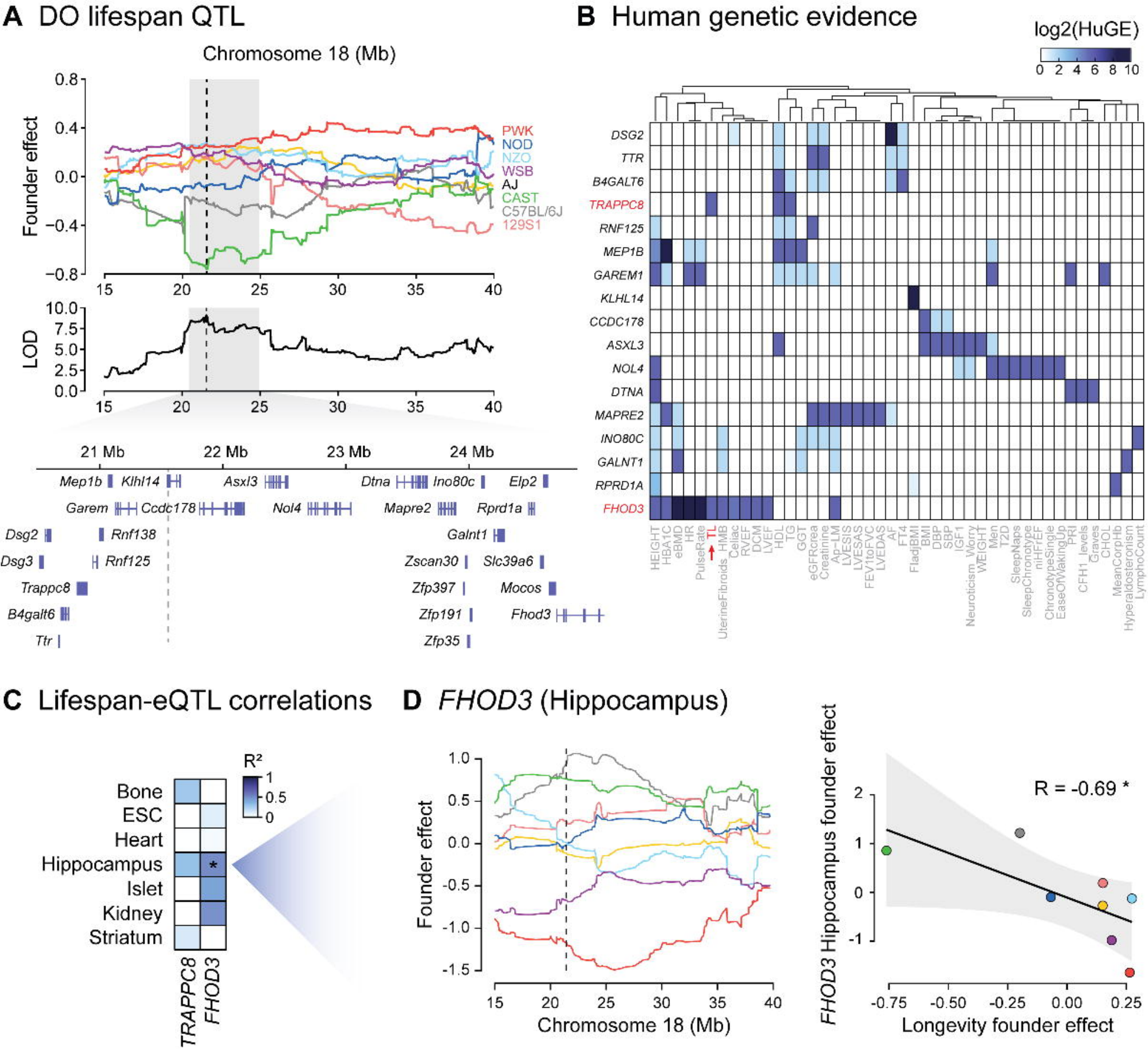
14. Bult, C.J., and Sternberg, P.W. (2023). The alliance of genome resources: transforming comparative genomics. Mammalian Genome *34*, 531. https://doi.org/10.1007/S00335-023-10015-2.
15. Pal, A., Barber, T.M., Van de Bunt, M., Rudge, S.A., Zhang, Q., Lachlan, K.L., Cooper, N.S., Linden, H., Levy, J.C., Wakelam, M.J.O., et al. (2012). PTEN Mutations as a Cause of Constitutive Insulin Sensitivity and Obesity . New England Journal of Medicine *367*, 1002–1011. https://doi.org/10.1056/NEJMOA1113966/SUPPL_FILE/NEJMOA1113966_DISCLOSURES.PDF.
16. Wijesekara, N., Konrad, D., Eweida, M., Jefferies, C., Liadis, N., Giacca, A., Crackower, M., Suzuki, A., Mak, T.W., Kahn, C.R., et al. (2005). Muscle-specific Pten deletion protects against insulin resistance and diabetes. Mol Cell Biol *25*, 1135–1145. https://doi.org/10.1128/MCB.25.3.1135-1145.2005.
17. McGarry, J.D., Leatherman, G.F., and Foster, D.W. (1978). Carnitine palmitoyltransferase I. The site of inhibition of hepatic fatty acid oxidation by malonyl-CoA. Journal of Biological Chemistry *253*, 4128–4136. https://doi.org/10.1016/S0021-9258(17)34693-8.
18. Schlaepfer, I.R., and Joshi, M. (2020). CPT1A-mediated Fat Oxidation, Mechanisms, and Therapeutic Potential. Endocrinology *161*. https://doi.org/10.1210/ENDOCR/BQZ046.
19. Kulkarni, P.P., Ekhlak, M., Singh, V., Kailashiya, V., Singh, N., and Dash, D. (2023). Fatty acid oxidation fuels agonist-induced platelet activation and thrombus formation: Targeting β-oxidation of fatty acids as an effective anti-platelet strategy. FASEB J *37*. https://doi.org/10.1096/FJ.202201321RR.
20. Kulkarni, P.P., Ekhlak, M., and Dash, D. (2023). Energy metabolism in platelets fuels thrombus formation: Halting the thrombosis engine with small-molecule modulators of platelet metabolism. Metabolism *145*, 155596. https://doi.org/10.1016/J.METABOL.2023.155596.
21. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559. https://doi.org/10.1186/1471-2105-9-559.
22. Li, H., Rukina, D., David, F.P.A., Yang Li, T., Oh, C.M., Gao, A.W., Katsyuba, E., Sleiman, M.B., Komljenovic, A., Huang, Q., et al. (2019). Identifying gene function and module connections by the integration of multispecies expression compendia. Genome Res *29*, 2034–2045. https://doi.org/10.1101/GR.251983.119/-/DC1.
23. Seldin, M.M., Koplev, S., Rajbhandari, P., Vergnes, L., Rosenberg, G.M., Meng, Y., Pan, C., Phuong, T.M.N., Gharakhanian, R., Che, N., et al. (2018). A Strategy for Discovery of Endocrine Interactions with Application to Whole-Body Metabolism. Cell Metab *27*, 1138-1155.e6. https://doi.org/10.1016/J.CMET.2018.03.015.
24. Priyadarshini, M., Kotlo, K.U., Dudeja, P.K., and Layden, B.T. (2018). Role of Short Chain Fatty Acid Receptors in Intestinal Physiology and Pathophysiology. Compr Physiol *8*, 1091. https://doi.org/10.1002/CPHY.C170050.
25. He, Y., Sun, M.M., Zhang, G.G., Yang, J., Chen, K.S., Xu, W.W., and Li, B. (2021). Targeting PI3K/Akt signal transduction for cancer therapy. Signal Transduction and Targeted Therapy 2021 6:1 *6*, 1–17. https://doi.org/10.1038/s41392-021-00828-5.
26. Shukla, S.D., Sergeant, S., and Kim, H.D. (1985). [3H]Inositol incorporation into phosphoinositides of pig reticulocytes. Biochim Biophys Acta *821*, 305–309. https://doi.org/10.1016/0005-2736(85)90100-2.
27. Masson, S.W.C., Cutler, H.B., and James, D.E. (2024). Unlocking metabolic insights with mouse genetic diversity. EMBO Journal. https://doi.org/10.1038/S44318-024-00221-2/ASSET/6DB0AFA2-B9A4-43F2-A270-BE11E295F103/ASSETS/GRAPHIC/44318_2024_221_FIG1_HTML.PNG.
28. Gasch, A.P., Payseur, B.A., and Pool, J.E. (2016). The Power of Natural Variation for Model Organism Biology. Trends Genet *32*, 147–154. https://doi.org/10.1016/J.TIG.2015.12.003.

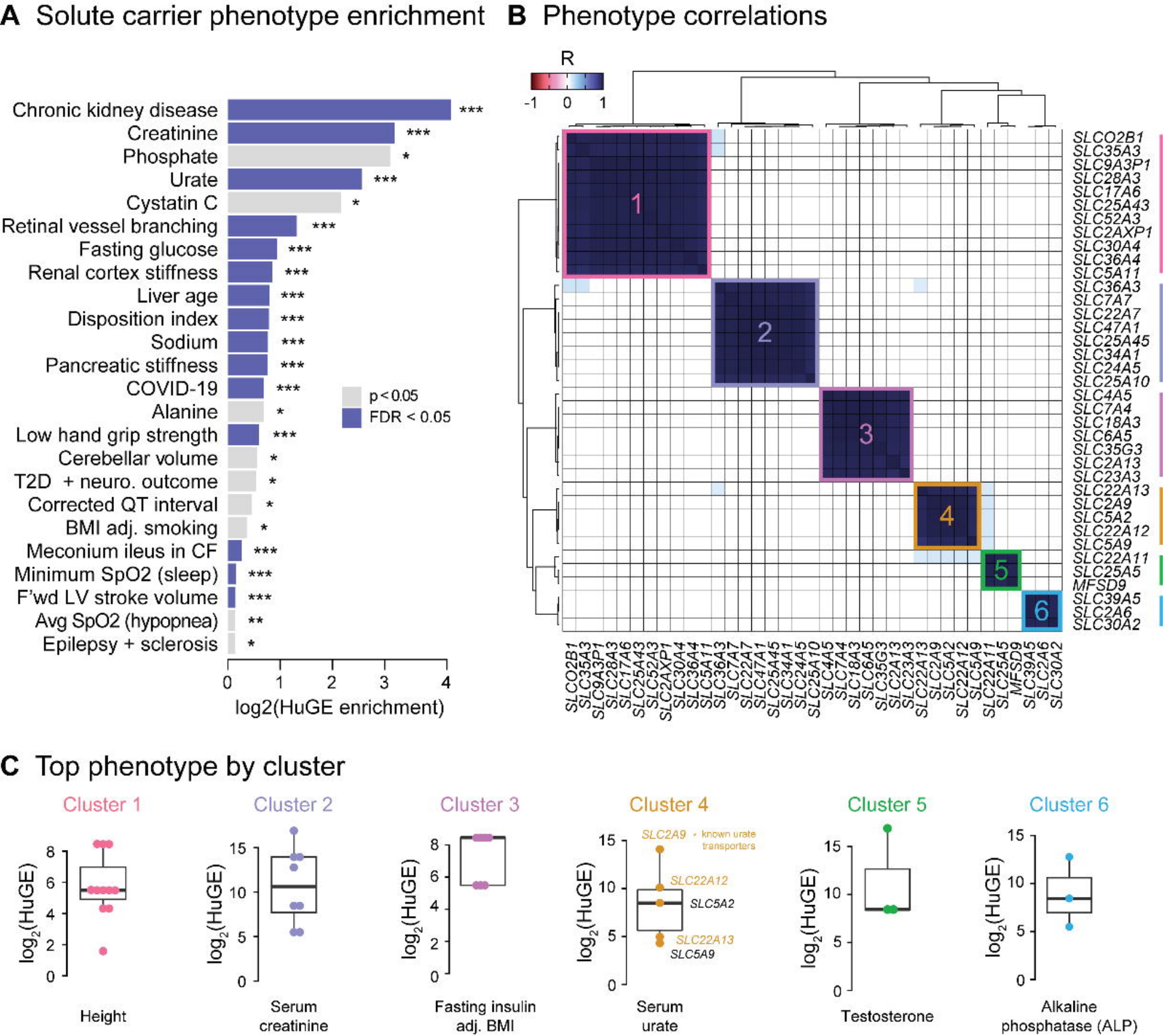
29. Di Francesco, A., Deighan, A.G., Litichevskiy, L., Chen, Z., Luciano, A., Robinson, L., Garland, G., Donato, H., Vincent, M., Schott, W., et al. (2024). Dietary restriction impacts health and lifespan of genetically diverse mice. Nature 2024 634:8034 *634*, 684–692. https://doi.org/10.1038/s41586-024-08026-3.
30. López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2023). Hallmarks of aging: An expanding universe. Cell *186*, 243–278. https://doi.org/10.1016/J.CELL.2022.11.001.
31. Lamb, C.A., Nühlen, S., Judith, D., Frith, D., Snijders, A.P., Behrends, C., and Tooze, S.A. (2016). TBC 1D14 regulates autophagy via the TRAPP complex and ATG 9 traffic . EMBO J *35*, 281–301. https://doi.org/10.15252/EMBJ.201592695/SUPPL_FILE/EMBJ201592695.REVIEWER_COMMENTS.PDF.
32. Zhang, Z., Bai, M., Barbosa, G.O., Chen, A., Wei, Y., Luo, S., Wang, X., Wang, B., Tsukui, T., Li, H., et al. (2020). Broadly conserved roles of TMEM131 family proteins in intracellular collagen assembly and secretory cargo trafficking. Sci Adv *6*, eaay7667. https://doi.org/10.1126/SCIADV.AAY7667.
33. Chadwick, S.R., and Lajoie, P. (2019). Endoplasmic Reticulum Stress Coping Mechanisms and Lifespan Regulation in Health and Diseases. Front Cell Dev Biol *7*, 84. https://doi.org/10.3389/FCELL.2019.00084.
34. Labunskyy, V.M., Gerashchenko, M. V., Delaney, J.R., Kaya, A., Kennedy, B.K., Kaeberlein, M., and Gladyshev, V.N. (2014). Lifespan Extension Conferred by Endoplasmic Reticulum Secretory Pathway Deficiency Requires Induction of the Unfolded Protein Response. PLoS Genet *10*, e1004019. https://doi.org/10.1371/JOURNAL.PGEN.1004019.
35. Aman, Y., Schmauck-Medina, T., Hansen, M., Morimoto, R.I., Simon, A.K., Bjedov, I., Palikaras, K., Simonsen, A., Johansen, T., Tavernarakis, N., et al. (2021). Autophagy in healthy aging and disease. Nature Aging 2021 1:8 *1*, 634–650. https://doi.org/10.1038/s43587-021-00098-4.
36. Hansen, M., Rubinsztein, D.C., and Walker, D.W. (2018). Autophagy as a promoter of longevity: insights from model organisms. Nature Reviews Molecular Cell Biology 2018 19:9 *19*, 579–593. https://doi.org/10.1038/s41580-018-0033-y.
37. Iskratsch, T., Lange, S., Dwyer, J., Kho, A.L., Dos Remedios, C., and Ehler, E. (2010). Formin follows function: a muscle-specific isoform of FHOD3 is regulated by CK2 phosphorylation and promotes myofibril maintenance. J Cell Biol *191*, 1159–1172. https://doi.org/10.1083/JCB.201005060.
38. Ushijima, T., Fujimoto, N., Matsuyama, S., Kan-O, M., Kiyonari, H., Shioi, G., Kage, Y., Yamasaki, S., Takeya, R., and Sumimoto, H. (2017). The actin-organizing formin protein Fhod3 is required for postnatal development and functional maintenance of the adult heart in mice. J Biol Chem *293*, 148. https://doi.org/10.1074/JBC.M117.813931.
39. Churchill, G.A., Gatti, D.M., Munger, S.C., and Svenson, K.L. (2012). The Diversity Outbred mouse population. Mamm Genome *23*, 713–718. https://doi.org/10.1007/s00335-012-9414-2.
40. Svenson, K.L., Gatti, D.M., Valdar, W., Welsh, C.E., Cheng, R., Chesler, E.J., Palmer, A.A., McMillan, L., and Churchill, G.A. (2012). High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics *190*, 437–447. https://doi.org/10.1534/GENETICS.111.132597.
41. Vincent, M., Gyuricza, I.G., Keele, G.R., Gatti, D.M., Keller, M.P., Broman, K.W., and Churchill, G.A. (2022). QTLViewer: an interactive webtool for genetic analysis in the Collaborative Cross and Diversity Outbred mouse populations. G3 Genes|Genomes|Genetics *12*. https://doi.org/10.1093/G3JOURNAL/JKAC146.
42. Stoeger, T., Gerlach, M., Morimoto, R.I., and Nunes Amaral, L.A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol *16*, e2006643. https://doi.org/10.1371/JOURNAL.PBIO.2006643.

43. Superti-Furga, G., Lackner, D., Wiedmer, T., Ingles-Prieto, A., Barbosa, B., Girardi, E., Goldmann, U., Gürtl, B., Klavins, K., Klimek, C., et al. (2020). The RESOLUTE consortium: unlocking SLC transporters for drug discovery. Nature Reviews Drug Discovery 2020 19:7 *19*, 429–430. https://doi.org/10.1038/d41573-020-00056-6.
44. Vitart, V., Rudan, I., Hayward, C., Gray, N.K., Floyd, J., Palmer, C.N.A., Knott, S.A., Kolcic, I., Polasek, O., Graessler, J., et al. (2008). SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat Genet *40*, 437–442. https://doi.org/10.1038/NG.106.
45. Dai, Y., and Lee, C.H. (2024). Transport mechanism and structural pharmacology of human urate transporter URAT1. Cell Research 2024 34:11 *34*, 776–787. https://doi.org/10.1038/s41422-024-01023-1.
46. Bahn, A., Hagos, Y., Reuter, S., Balen, D., Brzica, H., Krick, W., Burckhardt, B.C., and Sabolic´, I. (2008). Identification of a New Urate and High Affinity Nicotinate Transporter, hOAT10 (SLC22A13) *. https://doi.org/10.1074/jbc.M800737200.
47. Ferrannini, E., and Solini, A. (2012). SGLT2 inhibition in diabetes mellitus: rationale and clinical prospects. Nature Reviews Endocrinology 2012 8:8 *8*, 495–502. https://doi.org/10.1038/nrendo.2011.243.
48. Chao, E.C., and Henry, R.R. (2010). SGLT2 inhibition — a novel strategy for diabetes treatment. Nature Reviews Drug Discovery 2010 9:7 *9*, 551–559. https://doi.org/10.1038/nrd3180.
49. Inthasot, S., Vanderhulst, J., Janssens, P., Daele, S. Van, Hoof, E. Van, Kint, C., Iconaru, L., and de Filette, J. (2024). A novel heterozygous likely pathogenic SLC5A2 variant in a diabetic patient with glucosuria and aminoaciduria. Endocrinol Diabetes Metab Case Rep *2024*. https://doi.org/10.1530/EDM-24-0065.
50. Novikov, A., Fu, Y., Huang, W., Freeman, B., Patel, R., van Ginkel, C., Koepsell, H., Busslinger, M., Onishi, A., Nespoux, J., et al. (2019). SGLT2 inhibition and renal urate excretion: role of luminal glucose, GLUT9, and URAT1. Am J Physiol Renal Physiol *316*, F173–F185. https://doi.org/10.1152/AJPRENAL.00462.2018.
51. Dong, M., Chen, H., Wen, S., Yuan, Y., Yang, L., Xu, D., and Zhou, L. (2023). The Mechanism of Sodium-Glucose Cotransporter-2 Inhibitors in Reducing Uric Acid in Type 2 Diabetes Mellitus. Diabetes, Metabolic Syndrome and Obesity *16*, 437–445. https://doi.org/10.2147/DMSO.S399343.
52. Tazawa, S., Yamato, T., Fujikura, H., Hiratochi, M., Itoh, F., Tomae, M., Takemura, Y., Maruyama, H., Sugiyama, T., Wakamatsu, A., et al. (2005). SLC5A9/SGLT4, a new Na+-dependent glucose transporter, is an essential transporter for mannose, 1,5-anhydro-D-glucitol, and fructose. Life Sci *76*, 1039–1050. https://doi.org/10.1016/J.LFS.2004.10.016.
53. Sano, R., Shinozaki, Y., and Ohta, T. (2020). Sodium–glucose cotransporters: Functional properties and pharmaceutical potential. J Diabetes Investig *11*, 770. https://doi.org/10.1111/JDI.13255.
54. Johnson, R.J. (2023). Intestinal Hyperuricemia as a Driving Mechanism for CKD. American Journal of Kidney Diseases *81*, 127–130. https://doi.org/10.1053/j.ajkd.2022.08.001.
55. Allayee, H., Farber, C.R., Seldin, M.M., Williams, E.G., James, D.E., and Lusis, A.J. (2023). Systems genetics approaches for understanding complex traits with relevance for human disease. Elife *12*. https://doi.org/10.7554/ELIFE.91004.
56. Connally, N., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., Cotsapas, C., Cassa, C.A., and Sunyaev, S. (2022). The missing link between genetic association and regulatory function. Elife *11*, e74970. https://doi.org/10.7554/ELIFE.74970.
57. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., URL https://www.R-project.org/.

58. Ooms, J. (2024). A Simple and Robust JSON Parser and Generator for R [R package jsonlite version 1.8.9]. CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.PACKAGE.JSONLITE.
59. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/BIOINFORMATICS/BTW313.
60. Wickham, H. (2016). ggplot2. https://doi.org/10.1007/978-3-319-24277-4.
61. Sievert, C. (2021). Interactive Web-Based Data Visualization with R, Plotly, and Shiny. J R Stat Soc Ser A Stat Soc *184*, 1150–1150. https://doi.org/10.1111/RSSA.12692.


A Human Genetic Evidence
PPARG
log(HuGE score)
Very strong
Moderate
Rare variants
Common variants
Bayesian inference
HuGE score
Simple interpretation
B Current implementation
ASSOCIATION TO FUNCTION
KNOWLEDGE PORTAL
Manual querying
Search gene, variant, region, phenotype, or tissue
Wait for results to load
Genetic support workflow
1 Find significant variants
2 Explore common & rare effects
3 Investigate HuGE scores
Repeat (2-5 min/gene)
C High throughput workflow
SYNTENY
Experimental candidates
API
multi species support
1 Upload genes
2 Acquire data
3 Visualise interactively
D HuGE scores
log2(HuGE)
E Phenotype correlations
Correlation (R)
F Human phenotype enrichment
AMD progression
Omega-3 fatty acids
DHA
Late onset AD
Valine
Peptic Ulcers
TSH
AD
BMI 6 months
BMI 1-5 yrs
BMI 3 months
BMI 1 yr
Asc. aorta distensibility
Psychiatric
Chylomicrons
Glycoproteins
Fat intake adj. BMI
AMD
Long QT syndrome
p < 0.05
FDR < 0.05
log2(HuGE enrichment)
A2F: 1hr
Synteny: 1 min 20 sec
15 min (294 phenotypes)


A DO lifespan QTL
Chromosome 18 (Mb)
Founder effect
PWK
NOD
NZO
WSB
AJ
CAST
C57BL/6J
129S1
LOD
21 Mb
22 Mb
23 Mb
24 Mb
Mep1b
Klhl14
Asxl3
Dtna
Ino80c
Elp2
Garem
Ccdc178
Nol4
Mapre2
Rprd1a
Dsg2
Rnf138
Galnt1
Dsg3
Rnf125
Zscan30
Slc39a6
Trappc8
Zfp397
Mocos
B4galt6
Zfp191
Fhod3
Ttr
Zfp35
B Human genetic evidence
log2(HuGE)
DSG2
TTR
B4GALT6
TRAPPC8
RNF125
MEP1B
GAREM1
KLHL14
CCDC178
ASXL3
NOL4
DTNA
MAPRE2
INO80C
GALNT1
RPRD1A
FHOD3
HEIGHT
HBA1C
eBMD
HR
PulseRate
TL
UterineFibroids
HMB
Celiac
RVEF
DCM
LVEF
HDL
TG
GGT
eGFRcrea
Creatinine
Ap-LM
LVESIS
LVESAS
FEV1toFVC
LVEDAS
AF
FT4
FladjBMI
BMI
DBP
SBP
IGF1
Neuroticism_Worry
WEIGHT
Men
T2D
SleepNaps
SleepChronotype
niHFrEF
ChronotypeSingle
EaseOfWakingUp
PRI
CFH1_levels
Graves
CHOL
MeanCorpHb
Hyperaldosteronism
LymphoCount
C Lifespan-eQTL correlations
R²
Bone
ESC
Heart
Hippocampus
Islet
Kidney
Striatum
TRAPPC8
FHOD3
D FHOD3 (Hippocampus)
Founder effect
Chromosome 18 (Mb)
FHOD3 Hippocampus founder effect
R = -0.69 *
Longevity founder effect


A Solute carrier phenotype enrichment
Chronic kidney disease
Creatinine
Phosphate
Urate
Cystatin C
Retinal vessel branching
Fasting glucose
Renal cortex stiffness
Liver age
Disposition index
Sodium
Pancreatic stiffness
COVID-19
Alanine
Low hand grip strength
Cerebellar volume
T2D + neuro. outcome
Corrected QT interval
BMI adj. smoking
Meconium ileus in CF
Minimum SpO2 (sleep)
F'wd LV stroke volume
Avg SpO2 (hypopnea)
Epilepsy + sclerosis
p < 0.05
FDR < 0.05
log2(HuGE enrichment)
B Phenotype correlations
R
-1 0 1
SLCO2B1
SLC35A3
SLC9A3P1
SLC28A3
SLC17A6
SLC25A43
SLC52A3
SLC2AXP1
SLC30A4
SLC36A4
SLC5A11
SLC36A3
SLC7A7
SLC22A7
SLC47A1
SLC25A45
SLC34A1
SLC24A5
SLC25A10
SLC4A5
SLC7A4
SLC18A3
SLC6A5
SLC35G3
SLC2A13
SLC23A3
SLC22A13
SLC2A9
SLC5A2
SLC22A12
SLC5A9
SLC22A11
SLC25A5
MFSD9
SLC39A5
SLC2A6
SLC30A2
C Top phenotype by cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
log2(HuGE)
Height
Serum creatinine
Fasting insulin adj. BMI
Serum urate
Testosterone
Alkaline phosphatase (ALP)
SLC2A9
known urate transporters
SLC22A12
SLC5A2
SLC22A13
SLC5A9