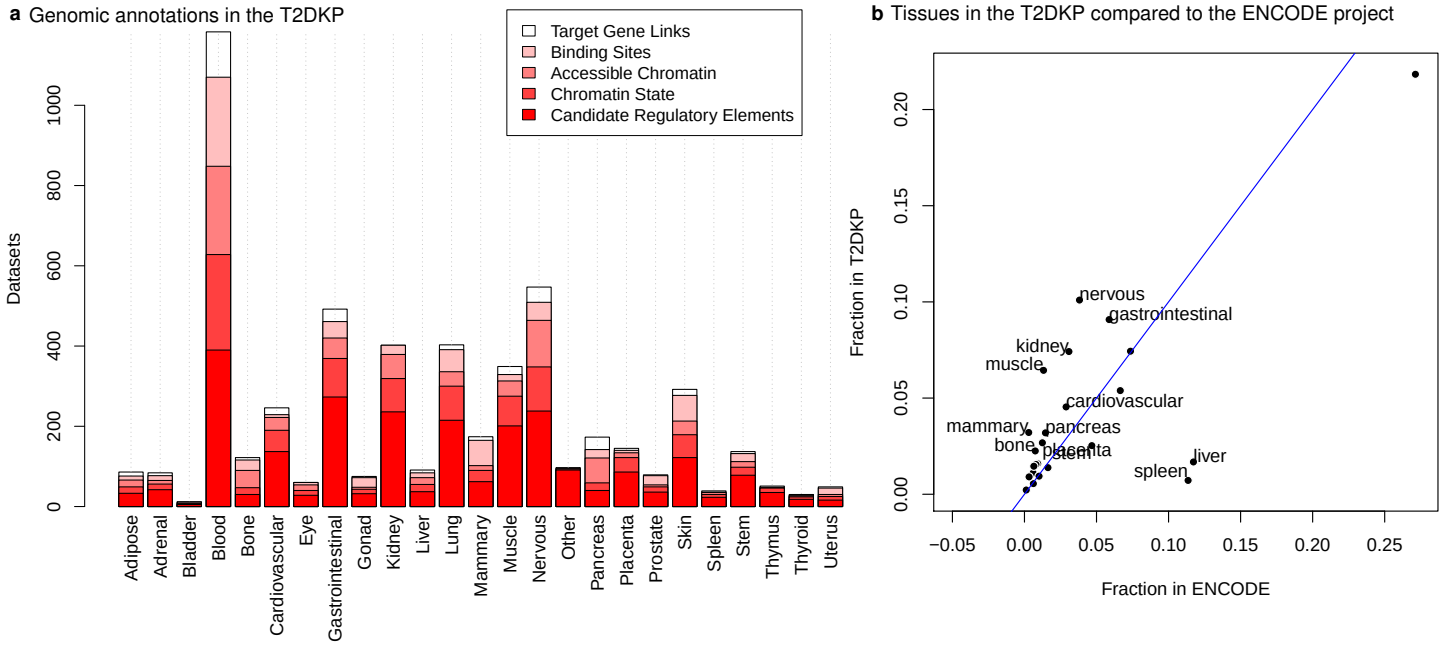
The Type 2 Diabetes Knowledge Portal: an open access genetic resource dedicated to type 2 diabetes and related traits

Supplementary Information

Contents

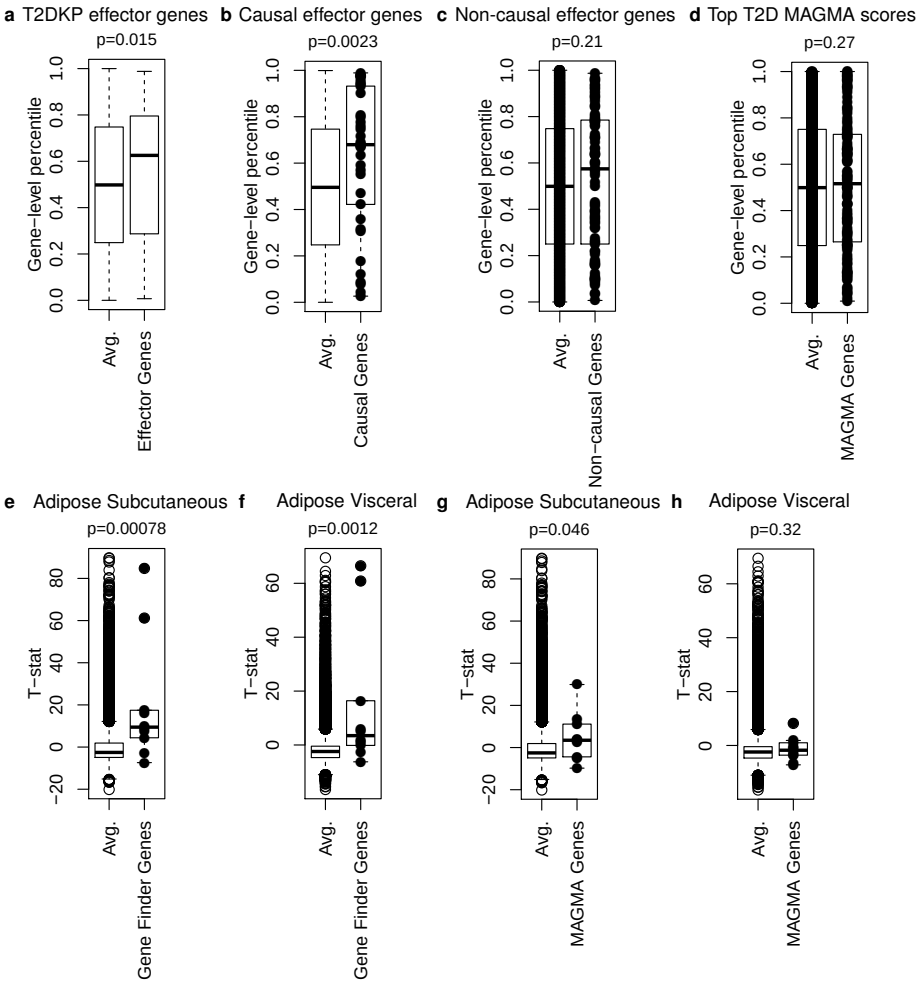
1 Supplementary Figures

Figure S1: **The T2DKP includes genomic annotations focused on diabetes-relevant tissues (Related to STAR Methods and Table S3).**



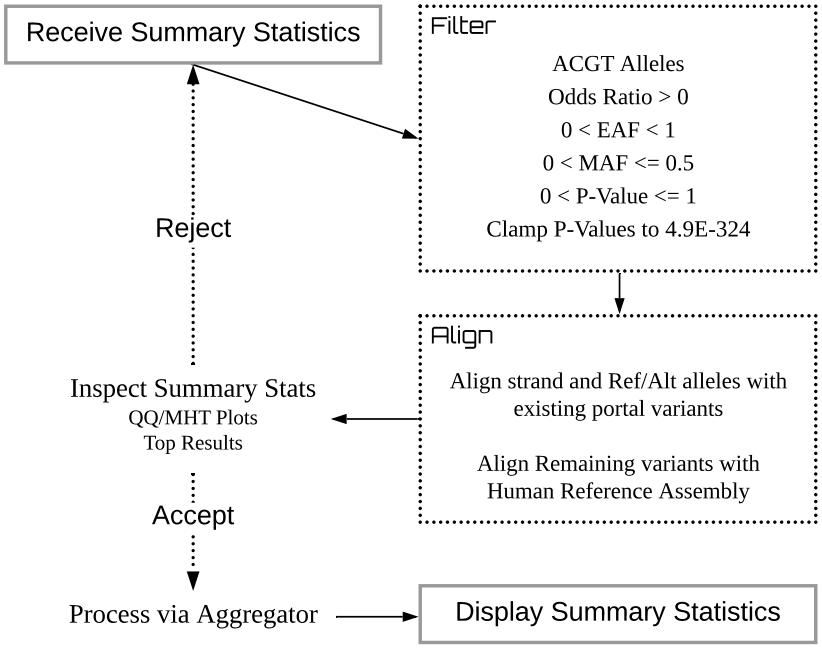
a. The T2DKP contains genomic annotations of accessible chromatin, transcription factor binding sites, candidate regulatory elements, and predicted chromatin state in 30 tissue categories. Raw datasets are stored in the Diabetes Epigenome Atlas where they are processed into tissue-level annotations. **b.** The distribution of genomic annotations in the T2DKP across tissues is similar to the distribution for genomic annotations in ENCODE, although the focus of the T2DKP on T2D-relevant tissues results in some tissues being over-represented. x-axis: fraction of datasets in ENCODE in each tissue. y-axis: fraction of datasets in the T2DKP in each tissue. labels: Tissues with a greater than 30% difference in proportion between the T2DKP and ENCODE.

Figure S2: **The T2DKP can be used to identify disease-relevant genes (Related to Figure 6 and STAR Methods).**



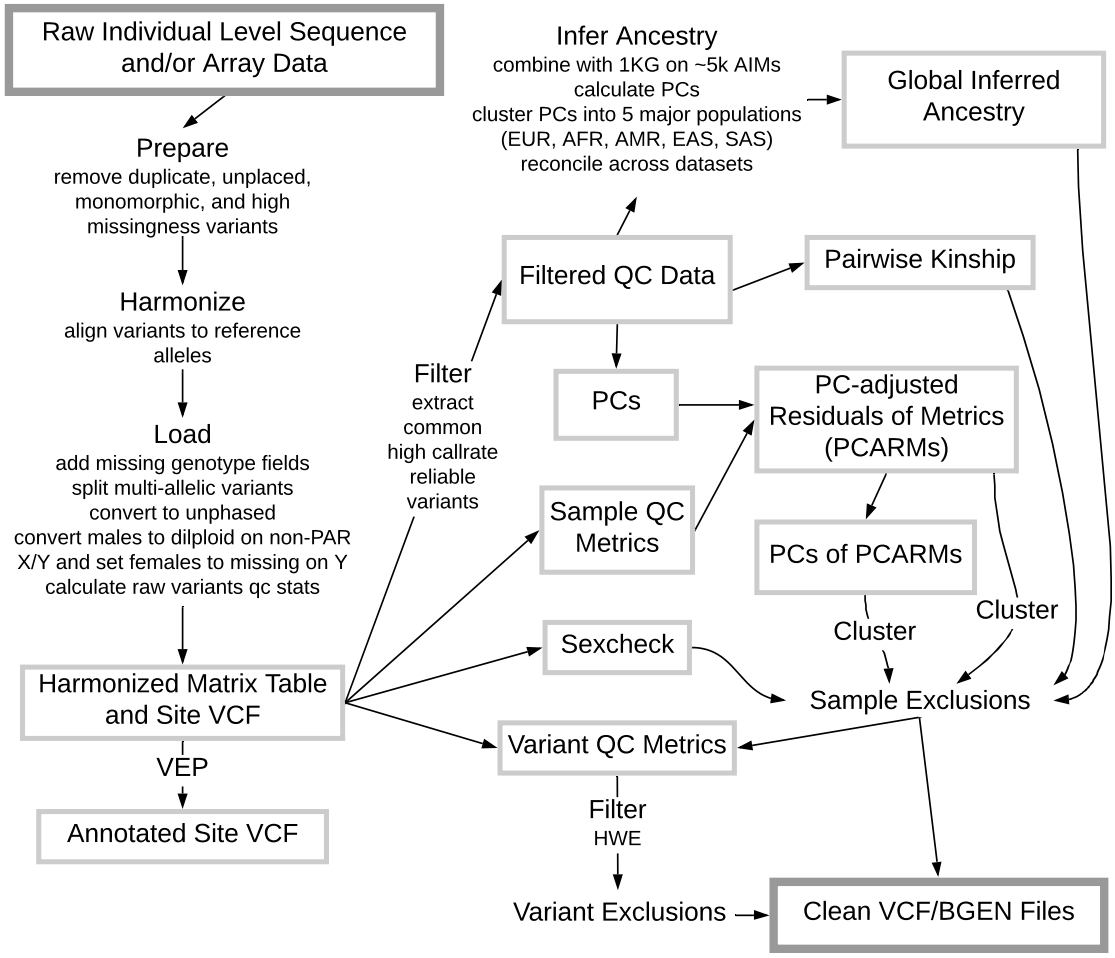
(a-d). We downloaded the list of 132 predicted effector genes from the T2DKP and used a two-sided Wilcox test to evaluate their enrichment for rare coding variant associations from the AMP-T2D-GENES study (**STAR Methods**); 128 of the effector genes were present in the analysis. Results are shown for **(a)** all genes on the effector gene list (n=128) **(b)** effector genes in the "Causal" category (n=38), **(c)** effector genes other than those in the "Causal" category (n=90), and **(d)** the top MAGMA genes in the portal (n=128). **(e-h)** We used the Gene Finder to identify 11 genes associated at $p < 2.5 \times 10^{-8}$ with an "insulin resistance signature" of FIadjBMI, TG, HDL, T2D, and WHRadjBMI. We then used a two-sided Wilcox test evaluate the enrichment of these genes for tissue-specific expression across 53 GTEx tissues, focusing on adipose tissue (a major site of insulin resistance). Results are shown for **(ef)** the 11 genes in **(e)** subcutaneous adipose tissue and **(f)** visceral adipose tissue, as well as **(gh)** the 11 genes with the strongest T2D MAGMA associations in **(g)** subcutaneous adipose tissue and **(h)** visceral adipose tissue. In all box plots, bold black lines represent medians and box boundaries represent first and third quartiles; whiskers extend to 1.5 times the interquartile range beyond the quartiles. Circles represent either points beyond the whiskers or show all data (if fewer than 50 data points were used for the box plot).

Figure S3: **Summary-level genetic data undergo a multi-step quality control procedure (Related to STAR Methods).**



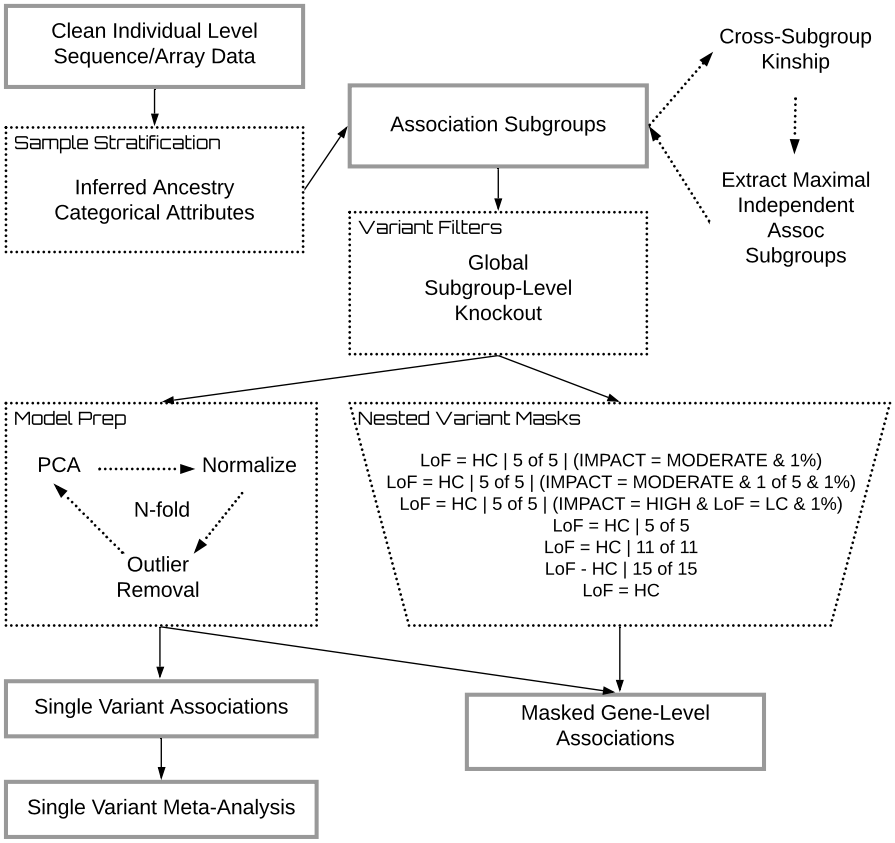
Each summary-level genetic dataset is first filtered to remove any variants with abnormal values for alleles, effect size, frequency, or p-values. Each variant's effect allele is then aligned to those of other datasets in the T2DKP (if the variant has been observed in another dataset) or to the non-reference allele (if it has not). QQ and Manhattan plots, as well as the top association results, are then inspected. If these plots look reasonable, and are approved by the dataset contributor, the dataset is advanced to the Data Aggregator.

Figure S4: **Individual-level genetic data are harmonized and subjected to quality control (Related to STAR Methods).**



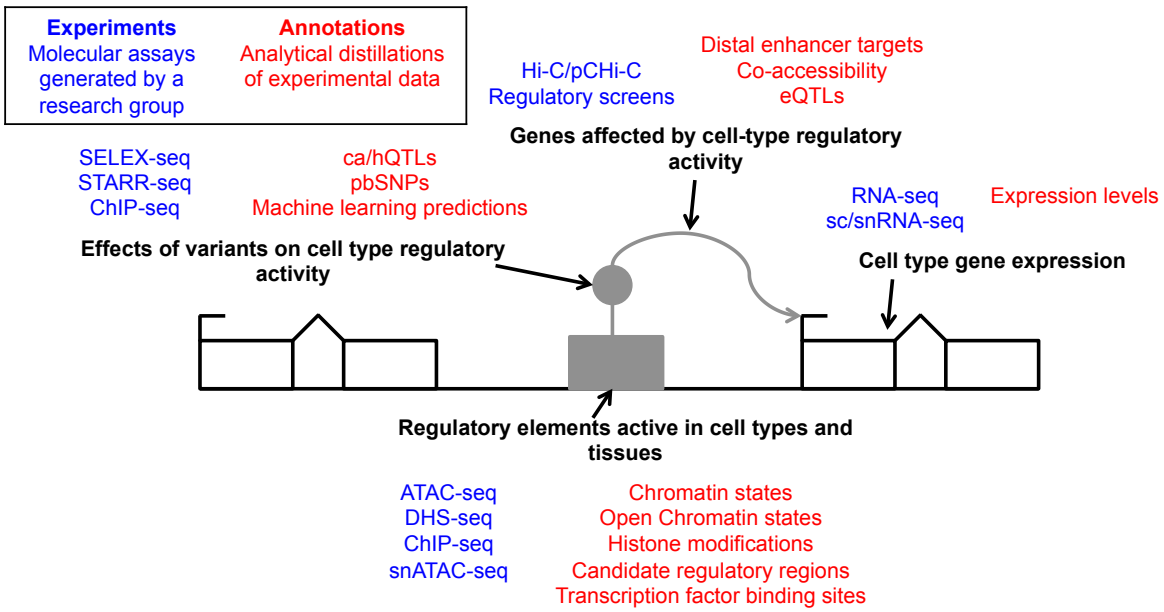
Individual-level genetic datasets include phenotypes and genotypes, often available across multiple SNP arrays. These data are subjected to a multi-step quality control process that harmonizes the genotyping data and then identifies outlier variants and samples (after controlling for inferred genetic ancestry). Sequence data undergo a similar quality control process.

Figure S5: **Individual-level genetic data are analyzed for single-variant and gene-level associations (Related to STAR Methods).**



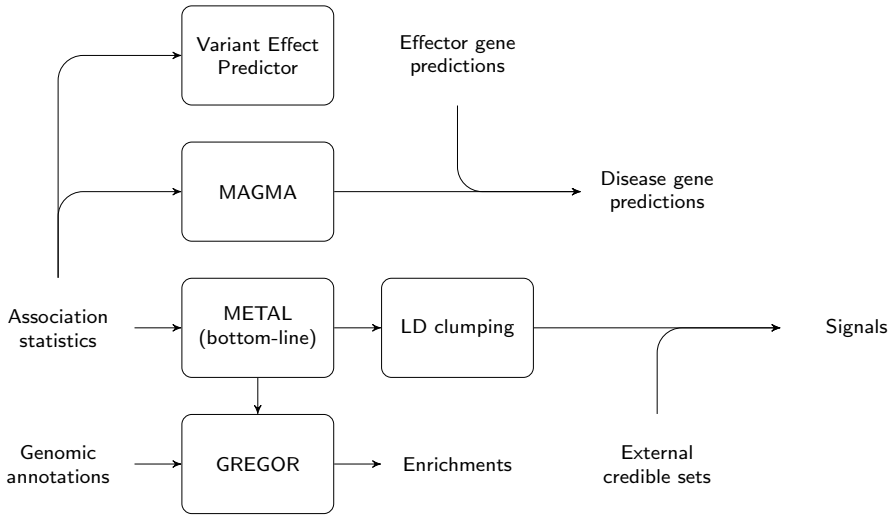
The first step in association analysis is to compute measures of genetic ancestry and relatedness. These are used to divide samples into subgroups, which are analyzed separately via single-variant analysis. Genetic ancestry, along with other potential confounders, are included as variables in the single-variant association test. Single-variant association statistics are then combined via meta-analysis. If sequence data are available (as opposed to SNP array data), gene-level associations are also computed across seven variant “masks”, defined based on variant bioinformatic annotations.

Figure S6: **The T2DKP stores four classes of genomic annotation (Related to STAR Methods).**



The T2DKP draws genomic annotations (red) from the Common Metabolic Diseases Epigenome Atlas, each generated by a different experimental procedure (blue). These genomic annotations are grouped into four broad classes (black), depending on the biological effect they predict.

Figure S7: **T2DKP datasets are processed through a series of bioinformatic methods (Related to STAR Methods).**



Association statistics and genomic annotations are analyzed by a series of bioinformatic methods to produce (for each trait in the T2DKP) predicted disease genes, independent association signals, and global enrichments of annotations for disease associations. Credible sets in some cases are contributed from external sources. Effector gene predictions are also contributed from external sources.

2 Consortia

The AMP-T2D Consortium. Gonçalo Abecasis, Beena Akolkar, Benjamin R. Alexander, Nicholette D. Allred, David Altshuler, Jennifer E. Below, Richard Bergman, Joline W.J. Beulens, John Blangero, Michael Boehnke, Krister Bokvist, Erwin Bottinger, Andrew P. Boughton, Donald Bowden, M Julia Brosnan, Christopher Brown, Kenneth Bruskiewicz, Noël P. Burtt, Mary Carmichael, Lizz Caulkins, Inês Cebola, John Chambers, Yii-Der Ida Chen, Andriy Cherkas, Audrey Y. Chu, Christopher Clark, Melina Claussnitzer, Maria C. Costanzo, Nancy J. Cox, Marcel den Hoed, Duc Dong, Marc Duby, Ravindranath Duggirala, Josée Dupuis, Petra J.M. Elders, Jesse M. Engreitz, Eric Fauman, Jorge Ferrer, Jason Flannick, Paul Flicek, Matthew Flickinger, Jose C. Florez, Caroline S. Fox, Timothy M. Frayling, Kelly A. Frazer, Kyle J. Gaulton, Clint Gilbert, Anna L. Gloyn, Todd Green, Craig L. Hanis, Robert Hanson, Andrew T. Hattersley, Quy Hoang, Hae Kyung Im, Sidra Iqbal, Suzanne B.R. Jacobs, Dong-Keun Jang, Tad Jordan, Tania Kamphaus, Fredrik Karpe, Thomas M. Keane, Seung K. Kim, Alexandria Kluge, Ryan Koesterer, Parul Kudtarkar, Kasper Lage, Leslie A. Lange, Mitchell Lazar, Donna Lehman, Ching-Ti Liu, Ruth J.F. Loos, Ronald Ching-wan Ma, Patrick MacDonald, Jeffrey Massung, Matthew T. Maurano, Mark I. McCarthy, Gil McVean, James B. Meigs, Josep M. Mercader, Melissa R. Miller, Braxton Mitchell, Karen L. Mohlke, Samuel Morabito, Claire Morgan, Shannon Mullican, Sharvari Narendra, Maggie C.Y. Ng, Lynette Nguyen, Colin N.A. Palmer, Stephen C.J. Parker, Antonio Parrado, Afshin Parsa, Aaron C. Pawlyk, Ewan R. Pearson, Andrew Plump, Michael Province, Thomas Quertermous, Susan Redline, Dermot F. Reilly, Bing Ren, Stephen S. Rich, J. Brent Richards, Jerome I. Rotter, Oliver Ruebenacker, Hartmut Ruetten, Rany M. Salem, Maike Sander, Michael Sanders, Dharambir Sanghera, Laura J. Scott, Sebanti Sengupta, David Siedzik, Xueling Sim, Xueling Sim, Preeti Singh, Robert Sladek, Kerrin Small, Philip Smith, Peter Stein, Dylan Spalding, Heather M. Stringham, Ying Sun, Katalin Susztak, Leen M. 't Hart, Daniel Taliun, Kent Taylor, Melissa K. Thomas, Jennifer A. Todd, Miriam S. Udler, Miriam S. Udler, Benjamin Voight, Marcin von Grotthuss, Andre Wan, Ryan P. Welch, David Wholley, Kaan Yuksel, Norann A. Zaghloul