

Effector Gene Prediction Symposium

AMP T2D Consortium

May 21, 2021



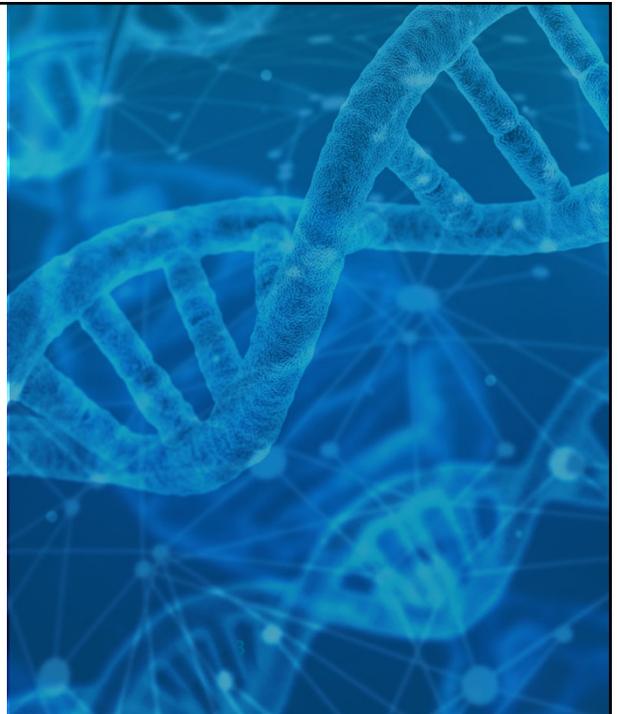
Agenda

- 11:00 a.m. 1. **Accelerating visualization and target prediction tools through AMP T2D**
Phil Smith and Melissa Thomas, AMP T2D Co-chairs
- 11:10 a.m. 2. **Considerations for target prioritization tools**
Eric Fauman, Symposium Chair
- 11:20 a.m. 3. **Target genes for T2D: a variant to causal gene approach—Effector Index**
Brent Richards
- 11:50 a.m. 4. **Connecting non-coding risk variants to target gene – ABC model**
Jesse Engreitz and Melina Claussnitzer
- 12:20 p.m. 5. **Strength of evidence score to predict effector genes – Heuristic model**
Anubha Mahajan and Mark McCarthy
- 12:50 p.m. 6. **Visualization of target prioritization tools on the AMP Common
Metabolic Diseases Knowledge Portal**
Noël Burt and Jason Flannick
- 1:20 p.m. 7. **PANEL DISCUSSION**
- 1:50 p.m. *Meeting Adjourn*



Accelerating visualization and target prediction tools through AMP T2D

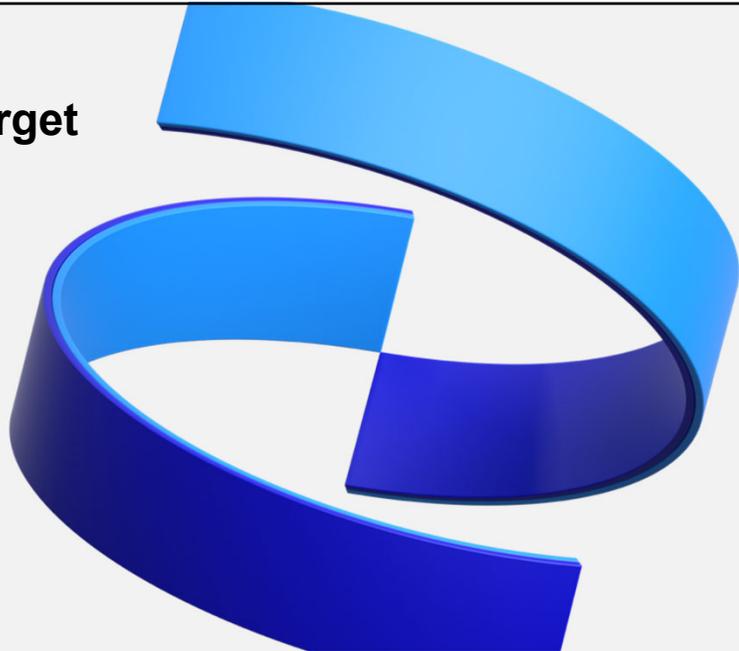
Phil Smith and Melissa Thomas
AMP T2D Co-Chairs



Considerations for target prioritization tools

Eric Fauman, PhD
Senior Scientific Director, Integrative
Biology, Internal Medicine Research Unit

May 21st, 2021

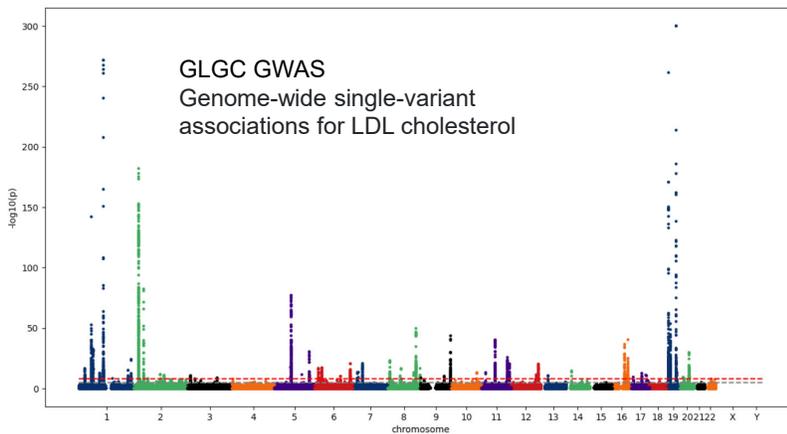


Breakthroughs that change patients' lives

4

4

We love to use human genetic evidence in drug discovery because it gives us greater confidence that the human gene is involved in the biological trait or disease we're interested in.



https://hugeamp.org/dinspector.html?dataset=GWAS_GLGC&phenotype=LDL

The first question human genetics can answer is "what"

What gene is contributing to this trait?

We're also interested in:

Why does this gene influence this trait?

and

How does the genetic variation impact the gene?



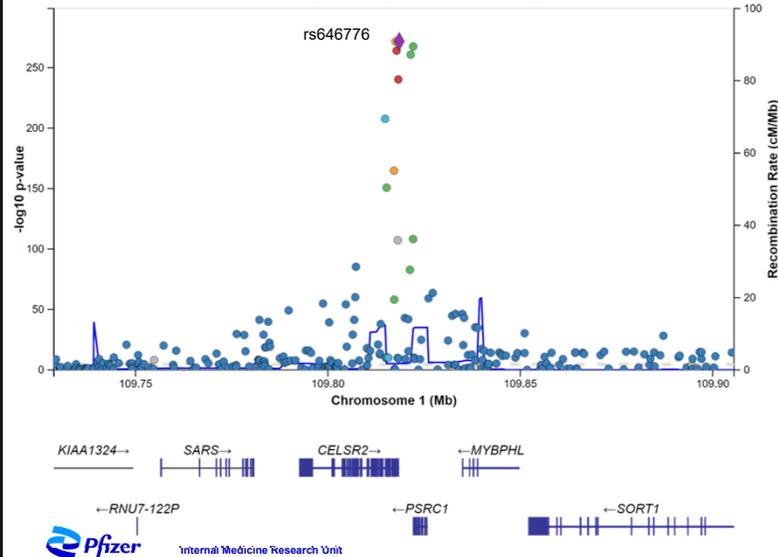
Internal Medicine Research Unit

5

5

GWAS is an amazing tool for establishing genetic associations for human traits
 But establishing the causal chain from variant to phenotype can be difficult

Willer CJ 2013 - Low-density lipoprotein (LDL) meta-analysis



Lead variant	rs646776
Causal or functional variant	?
Impact of variant on effector transcript	?
Causal gene/effector transcript	?
Function or activity of effector transcript	?
Measured phenotype	LDL-C levels

6

The full causal path from observed lead variant to observed phenotype goes through many steps, which can be evaluated somewhat independently

Lead variant	rs646776
Causal or functional variant	rs12740374
Impact of variant on effector transcript	C/EBP TF binding site altering the liver expression of SORT1
Causal gene/effector transcript	SORT1
Function or activity of effector transcript	Liver VLDL secretion
Measured phenotype	LDL-C levels

NIH Public Access
 Author Manuscript
 Published in final edited form as:
 Nature. 2010 August 5; 466(7307): 714-719. doi:10.1038/nature09266.

From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus

Kiran Musunuru^{1,2,3,14}, Alanna Strong^{4,14}, Maria Frank-Kamenetsky⁵, Noemi E. Lee¹, Tim Ahfeldt¹, Katherine V. Zecher¹, Xiaoyu Li¹, Jui Li¹, Nicolas Kuperwasser¹, Vera M. Rada¹, James J. Pirovano¹, Brian Muchmore¹, Ludmila Prokushina-Ossion¹, Jennifer L. Hall¹, Eric E. Schaaf⁶, Carlos R. Morales¹⁰, Sissel Lund-Katz¹¹, Michael C. Phillips¹¹, Jamie Wong¹², William Castelli¹², Timothy Ralston¹³, Kenneth G. Spivey¹³, Mary Ordo-Melander⁷, Ole Melander⁷, Victor Kotelnitsky⁸, Kevin Fitzgerald⁸, Ronald M. Krauss¹³, Chad A. Cowan¹⁵, Sakar Kothiresan^{16,17}, and Daniel J. Rader^{4,15}

¹Cardiovascular Research Center and Center for Human Genetic Research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA
²Broad Institute, Cambridge, Massachusetts 02142, USA
³Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA
⁴Institute for Translational Medicine and Therapeutics, Institute for Diabetes, Obesity and Metabolism, and Cardiovascular Institute, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA
⁵Nyram Pharmaceuticals, Inc., Cantonville, Massachusetts 02142, USA
⁶Department of Biochemistry and Molecular Biology II, Molecular Cell Biology, University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany
⁷Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
⁸Program in Cardiovascular Translational Genomics, Lilliehall Heart Institute, University of Minnesota, Minneapolis, Minnesota 55455, USA
⁹Sage Bionetworks, Seattle, Washington 98109, USA
¹⁰Department of Anatomy and Cell Biology, McGill University, Montreal, Quebec H3A 2B2, Canada
¹¹The Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

Musunuru et al, Nature. 2010 466(7307):714-719

7

— The goal of Effector Gene Prediction is to identify or rank the most likely causal gene at a locus

Two fundamental approaches

Lead variant	rs646776
Causal or functional variant	?
Impact of variant on effector transcript	?
Causal gene/effector transcript	?
Function or activity of effector transcript	?
Measured phenotype	LDL-C levels

Genomic features, independent of biological function of the genes:

- Proximity
- Sequence consequence
- eQTL/pQTL colocalization
- Chromatin conformation
- Co-accessibility
- Perturbation experiments
- etc

Biological features, independent of genomic context:

- Known biological function or pathways
- Mouse K/O
- Rare diseases
- Known drug targets
- etc

8

— The goal of Effector Gene Prediction is to identify or rank the most likely causal gene at a locus

Lead variant	rs646776
Causal or functional variant	?
Impact of variant on effector transcript	?
Causal gene/effector transcript	?
Function or activity of effector transcript	?
Measured phenotype	LDL-C levels

How does the genetic variation impact the gene?
Contributes to our understanding of “horsepower”

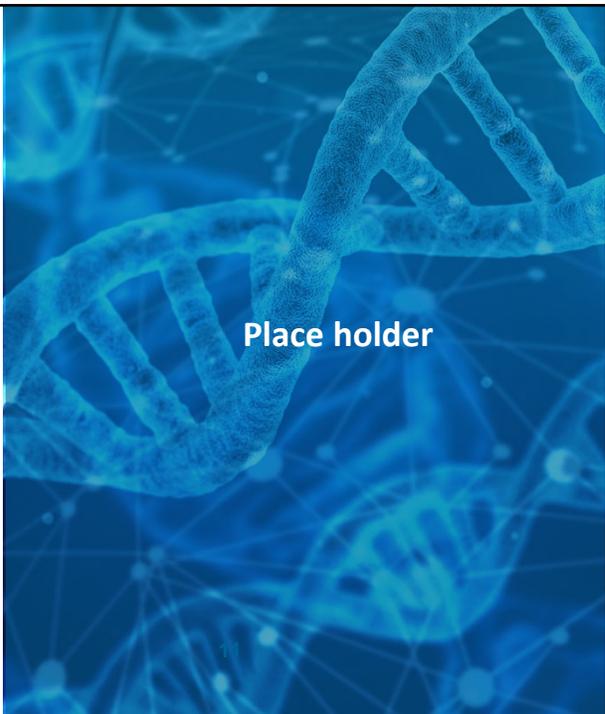
What gene is contributing to this trait?
Which gene should we go after?

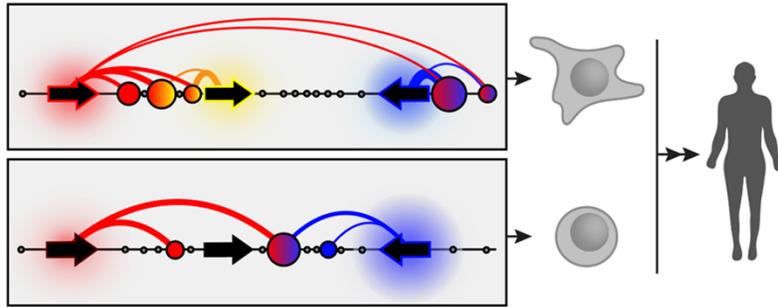
Why does this gene influence this trait?
What pathway or mechanism should we be targeting?

9

Target genes for T2D: a variant to causal gene approach– Effector Index

Brent Richards





Genome-wide maps of enhancer regulation connect risk variants to disease genes

Melina Claussnitzer, Ph.D.

Broad Institute of MIT and Harvard
Department of Medicine, Beth Israel Deaconess
Medical Center Harvard Medical School

Jesse Engreitz, Ph.D.

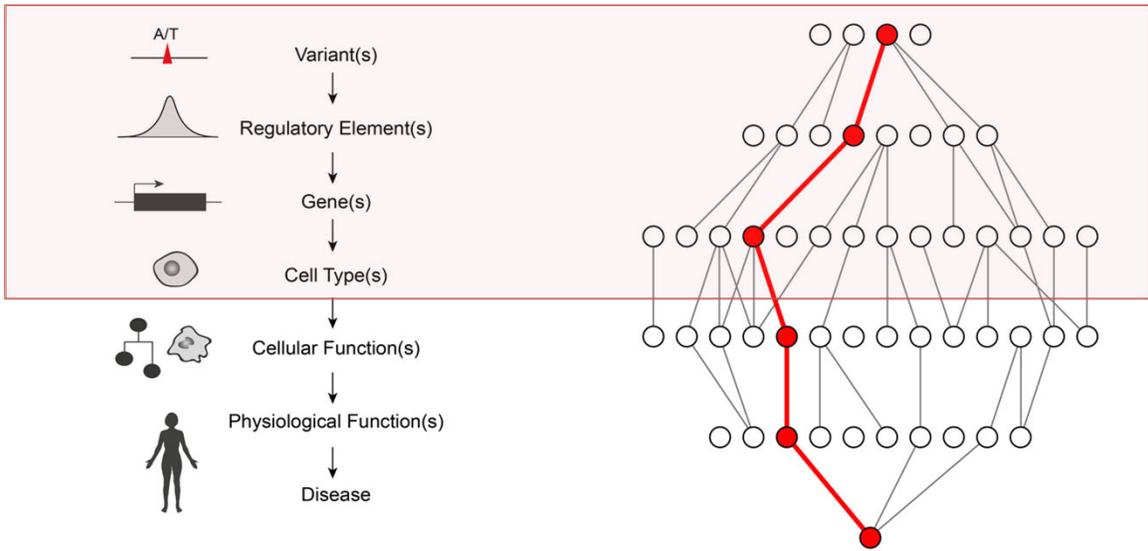
BASE Initiative, Moore Children's Heart Center
Department of Genetics, Stanford University
Broad Institute of MIT and Harvard

12

No disclosures to report

13

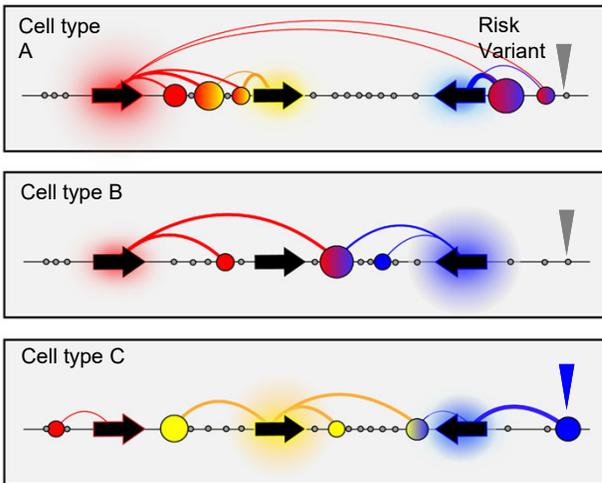
Connecting all T2D variants to enhancers and cell types



14

Build a Human Gene Regulation Map:

Which variants and enhancers regulate which genes in which cell types?



Challenge:

21,000 genes
 ×
 1,000,000s of enhancers
 ×
 1000s of cell types

15

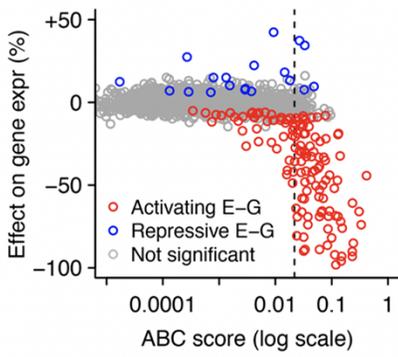
A simple formula predicts E-G connections

Predicted effect of enhancer on gene = $\frac{\text{Activity} \times \text{Contact}}{\sum (\text{Activity} \times \text{Contact})}$

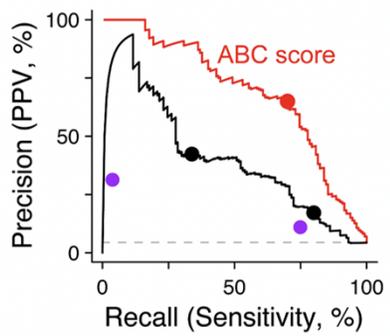
$\text{Activity} = \text{DHS} \times \text{H3K27ac}$
 $\text{Contact} = \text{Hi-C contact frequency}$



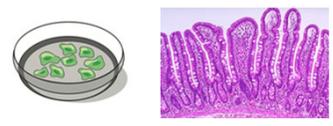
Predict quantitative effects:



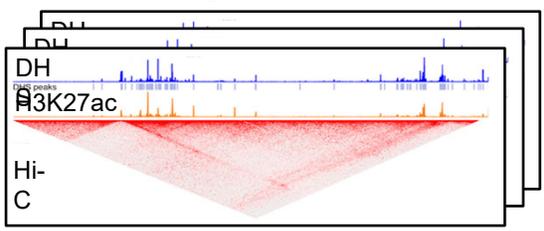
Classify regulatory E-G pairs:



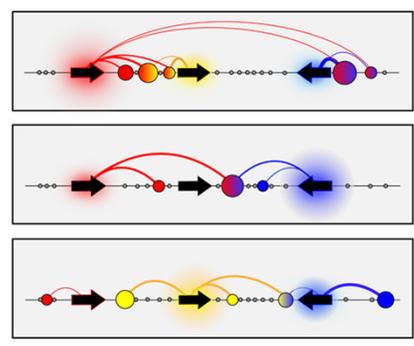
ABC predicts enhancer-gene connections across cell types



Epigenome maps



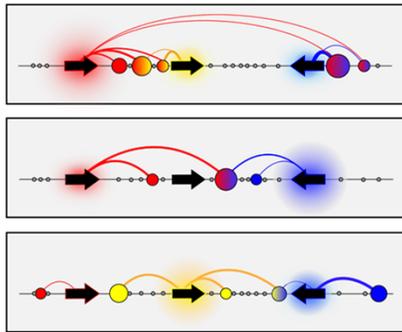
ABC



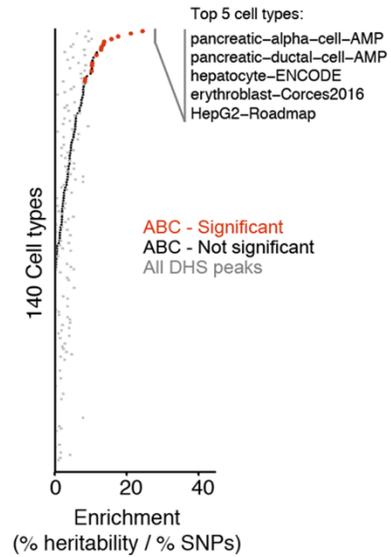
Enhancer-promoter maps

We can now create draft regulatory maps of the noncoding genome from epigenomic data across many cell types

How well does ABC link GWAS variants to known **cell types**?



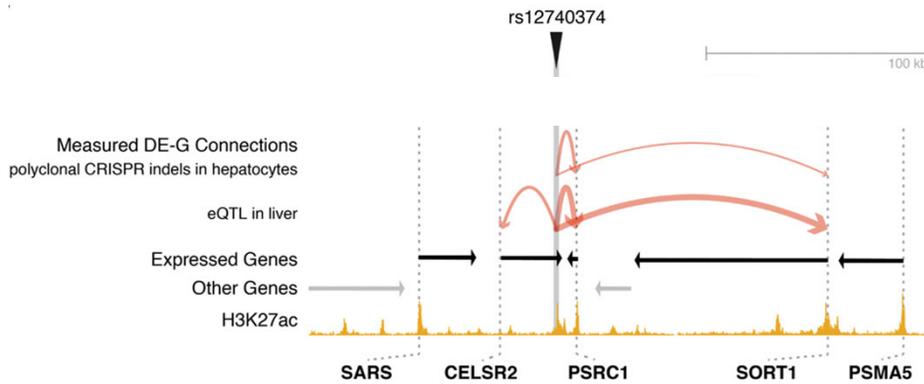
AMP-T2D:
Pancreatic cells, adipocytes,
hepatocytes, muscle



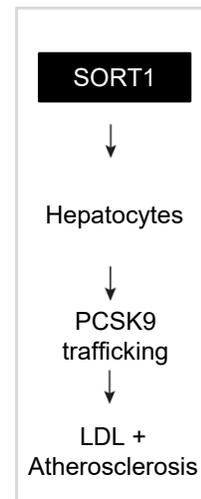
Melina Claussnitzer, Alisa Manning, Tim Majarian

18

Does ABC link variants to **known genes**?



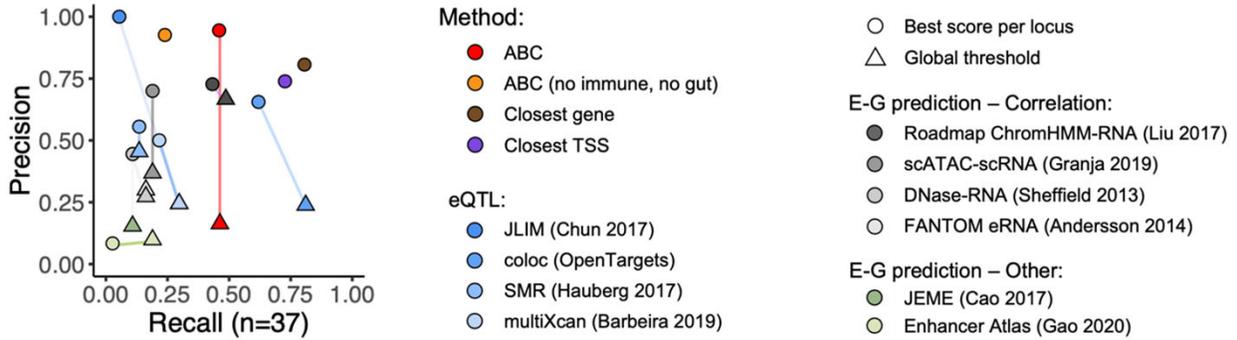
Musunuru *et al.*
Nature 2010



Fulco *et al.* Nat Genet 2019

19

Set 1: IBD genes



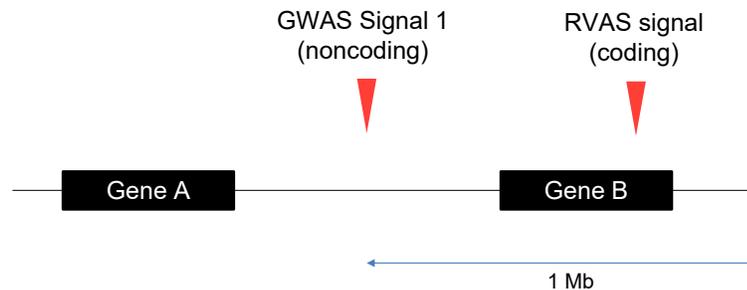
IBD Known Genes: Curated by domain expert based on prior evidence from coding variation, knowledge of gene function, mouse models

Nasser *et al.* Nature 2021

20

Can we collect other sets of “silver-standard” genes?

Idea: Use genes containing rare coding variants as positive controls

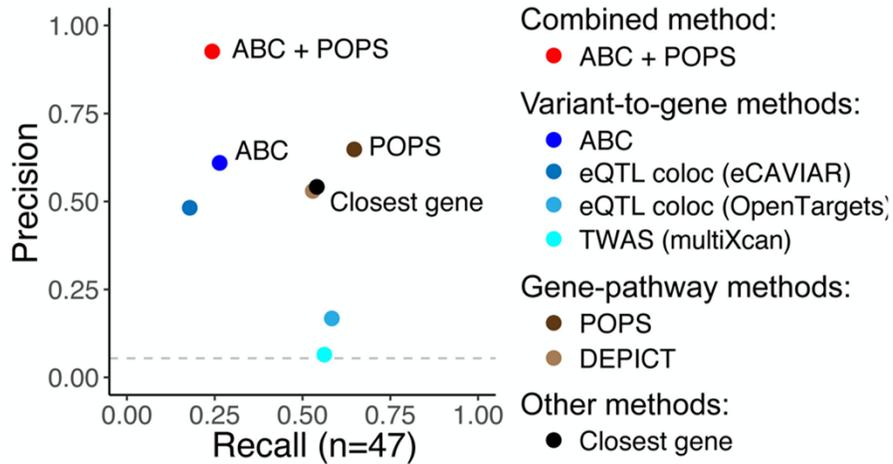


with Elle Weeks, Hilary Finucane

21

Set 2: LDL cholesterol

LDL cholesterol: Genes carrying rare coding variants from RVAS

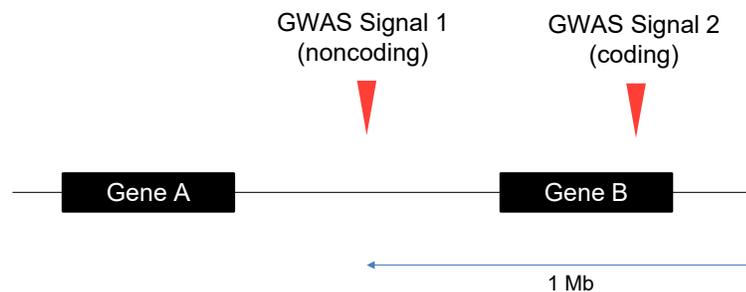


with Hilary Finucane, Elle Weeks

22

Can we collect a larger set of “silver-standard” genes?

Idea: Use genes containing coding variants as positive controls



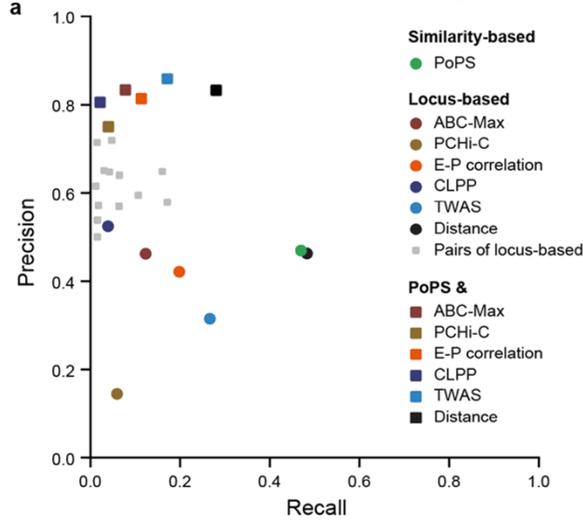
There are >600 such examples across 73 UK Biobank traits

Tejal Patwardhan, Elle Weeks, Jacob Ulirsch, Hilary Finucane

23

Set 3. 73 traits in UK Biobank

~600 genes with nearby common coding variants across 73 traits

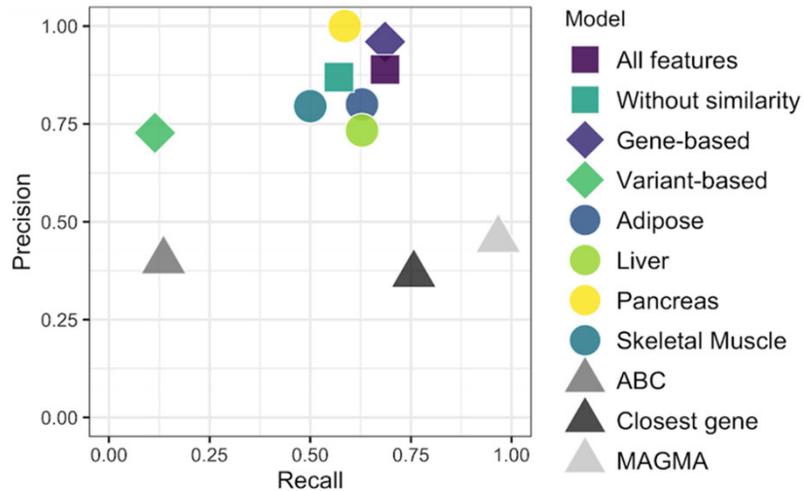


Weeks *et al.*, medRxiv 2020

24

Set 4. Type 2 diabetes

“Causal” effector genes from T2DKP



Tim Majarian, Alisa Manning

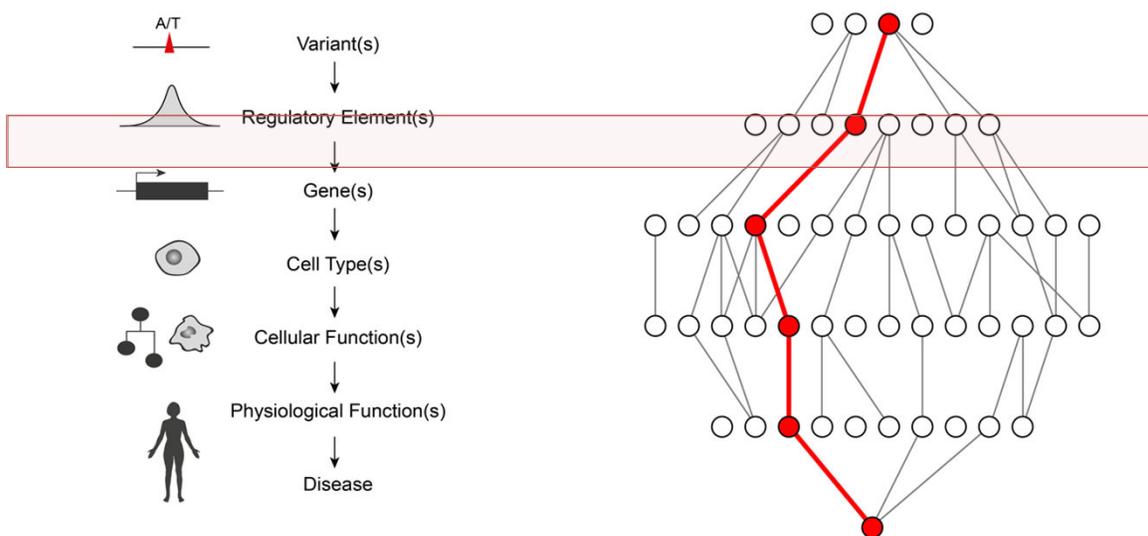
25

Considerations for gene prediction

- ABC performs well at identifying known genes and cell types compared to other regulatory “bottom-up” approaches
- Enhancers often regulate multiple genes → links GWAS variants to multiple candidates
- Combining “bottom-up” and “top-down” approaches likely to be needed to uniquely identify causal genes
- Considerations for comparing across traits:
 - How is the “silver-standard” gene set defined?
 - How many GWAS signals are there for this trait?
 - What statistical fine-mapping method was used?
 - How well are the relevant cell types covered in our ABC datasets?

26

Connecting all T2D variants to enhancers and cell types

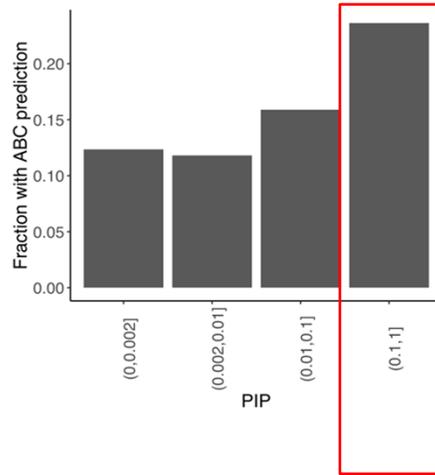
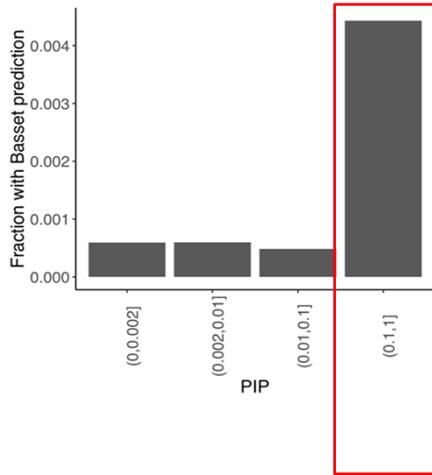


27

Connecting risk variants to enhancers to genes

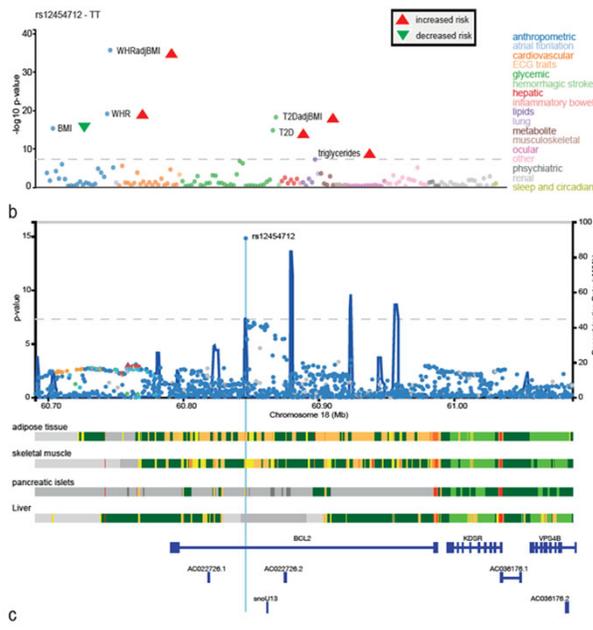
CNN variant predictions (Basset score ≥ 0.03)

ABC V-G predictions

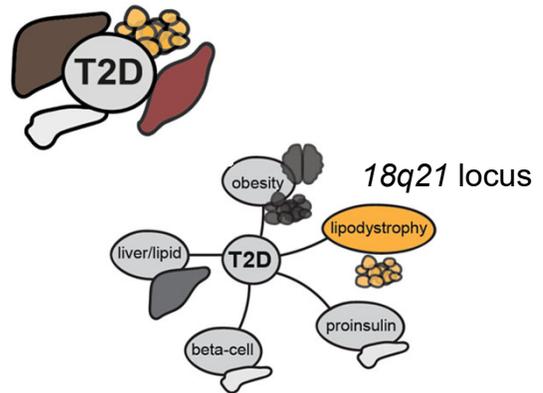


30

Connecting V2F for the 18q21 metabolic risk locus



Metabolically unhealthy lean phenotype

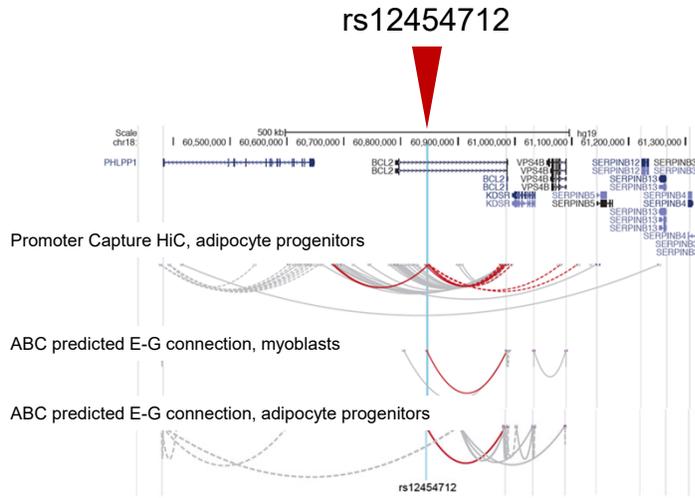


Udler et al, 2018

31

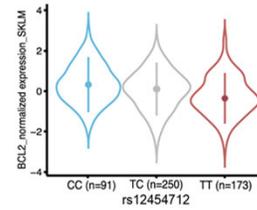
Connecting V - E – G for the 18q21 metabolic risk locus

ABC effector gene predictions

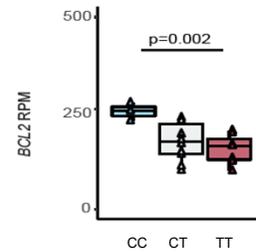


eQTLs

Skeletal muscle (STARNET database)



Adipocyte progenitors

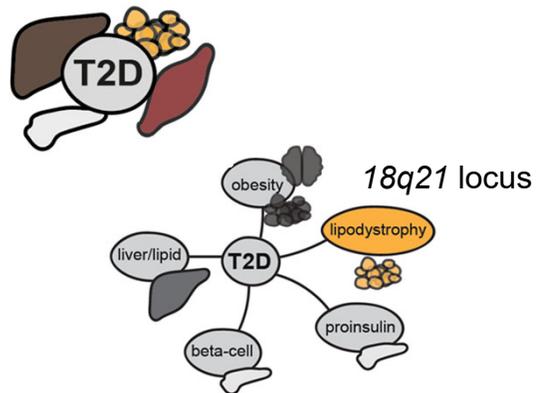
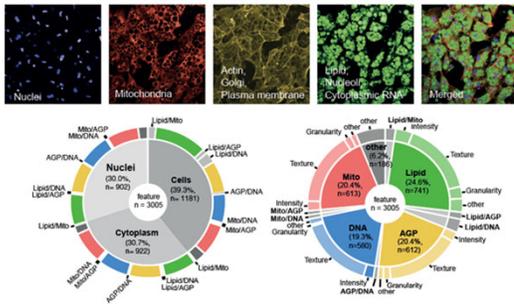


32

Connecting V2F for the 18q21 metabolic risk locus

Are the predicted variants and genes affecting *disease-relevant cellular programs*?

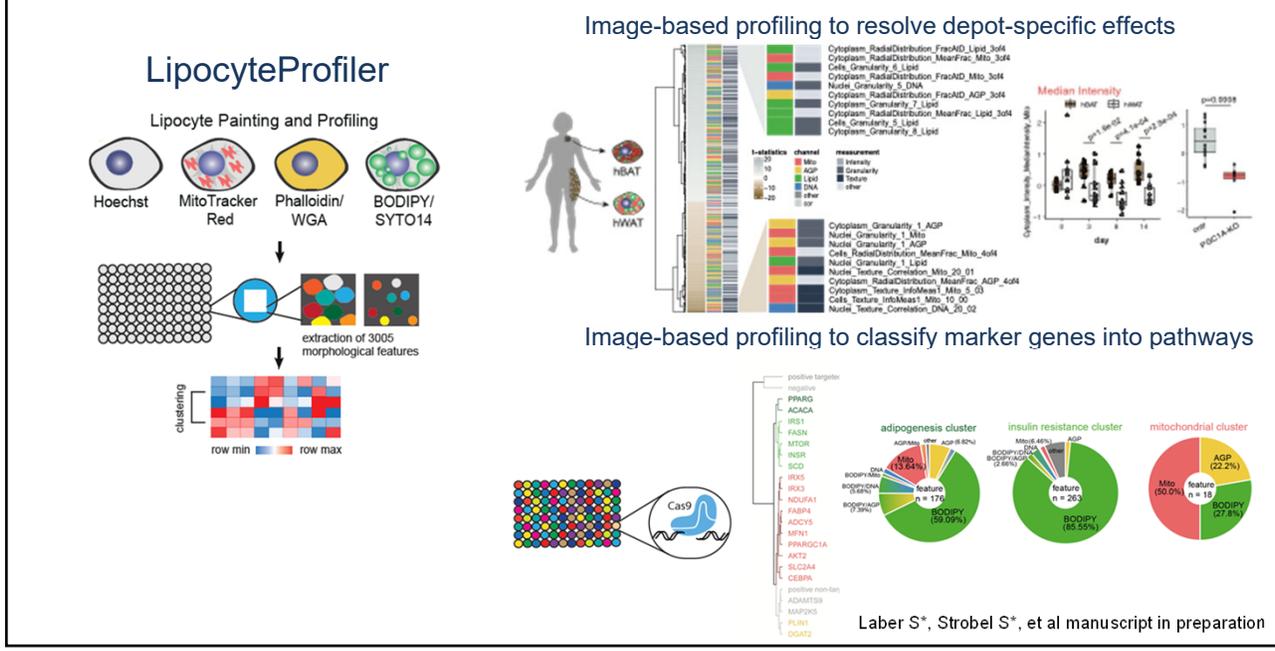
High-dimensional image-based profiling by LipocyteProfiler



Laber S*, Strobel S*, et al manuscript in preparation

33

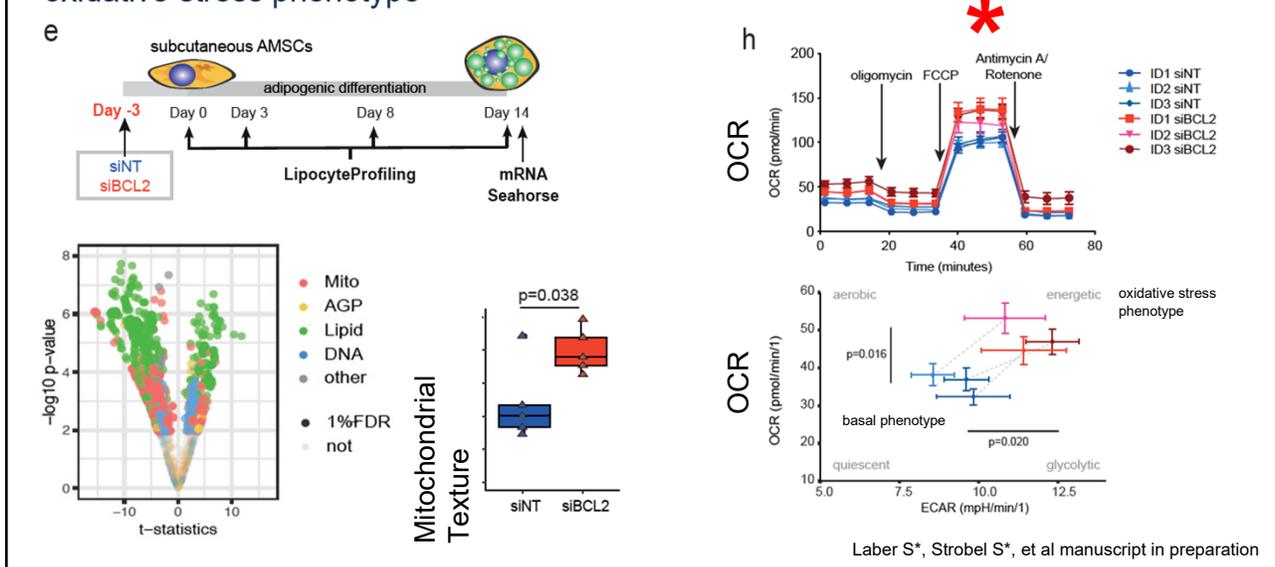
Connecting V2F for the 18q21 metabolic risk locus



34

Connecting G2F for the 18q21 metabolic risk locus

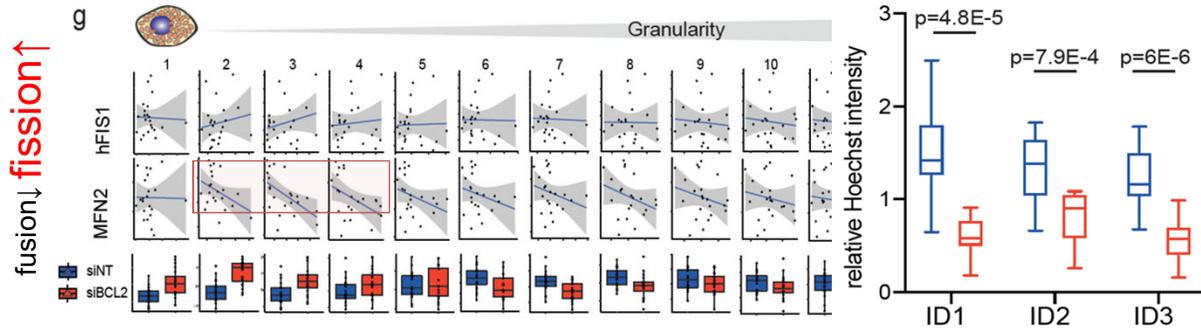
Directed *BCL2* perturbation in subcutaneous adipocyte progenitors results in an oxidative stress phenotype



35

Connecting G2F for the 18q21 metabolic risk locus

Directed *BCL2* perturbation in subcutaneous adipocyte progenitors results in increased mitochondrial fission and apoptosis

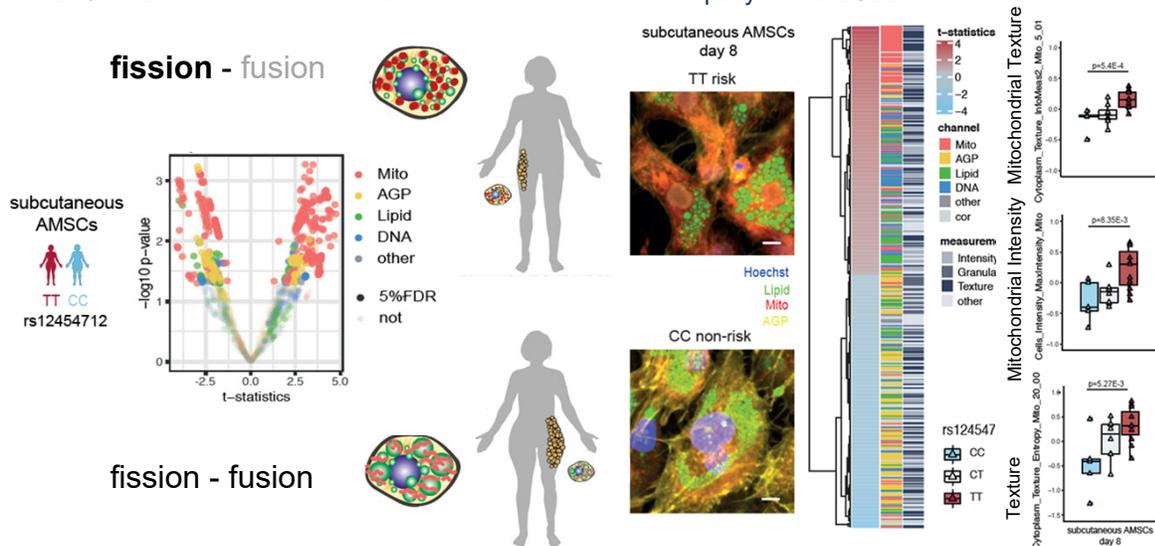


Laber S*, Strobel S*, et al manuscript in preparation

36

Connecting V2F for the 18q21 metabolic risk locus

rs12454712 controls mitochondrial fission in subcutaneous adipocytes via *BCL2*



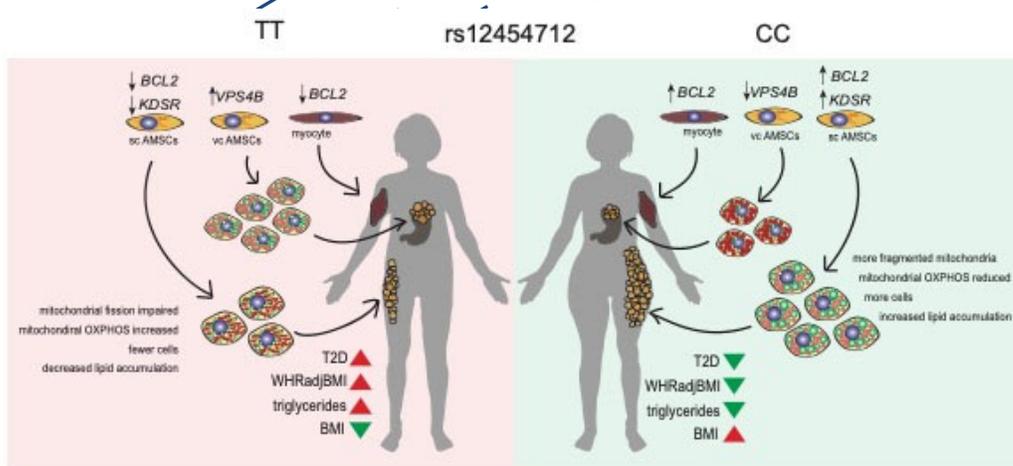
Laber S*, Strobel S*, et al manuscript in preparation

37

Connecting V2F for the *18q21* metabolic risk locus

rs12454712 affects multiple genes in multiple cell types, e.g. *VPS4B* in visceral adipocytes

cell context visceral adipocytes



38

Next steps

Bottom-Up (Regulatory models)

Epigenomic data in *all* disease relevant cell states

Deep eQTL single-cell datasets in right cell types and states

Better computational inferences of variant function

Build ABC all of the “right” cell types

Better definitions of cellular programs

Multi-modal single-cell data in cell types and cell contexts across many individuals (scRNA-seq, scATAC-seq, high-dimensional image-based profiling, etc);

Large-scale reverse genetics screens with single cell read-outs (Perturb-seq, Pooled optical screens, etc)

39

Acknowledgements

Claussnitzer Lab:

Nasa Sinnott-Armstrong
 Giacomo Deodato
 Samantha Laber
 Sophie Strobel
 Hesam Dashti

AMP Collaborators:

Tim Majarian
 Alissa Manning
 Jose Florez
 Kyle Gaulton
 Mark McCarthy
 Eric Fauman

T2DKP Team:

Jason Flannick
 Noel Burt
 Ben Alexander



Engreitz/Lander Labs:

Joe Nasser
 Charlie Fulco
 Drew Bergman
 Philine Guckelberger
 Ray Jones
 Tejal Patwardhan
 Tung Nguyen
 Ben Doughty
 Glen Munson
 Michael Kane
 Kate Lawrence
 Vidya Subramanian
 Kaite Zhang
 Shari Grossman
 Brian Cleary
 Ryan Collins



Hilary Finucane

Ran Cui
 Elle Weeks
 Jacob Ulirsch
 Masahiro Kanai
 Nir Hacohen
 John Ray
 Ang Cui
 Tom Eisenhaure
 Larry Schweitzer
 Matteo Gentili

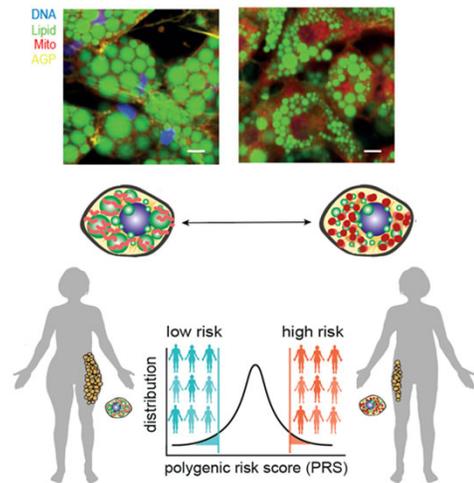
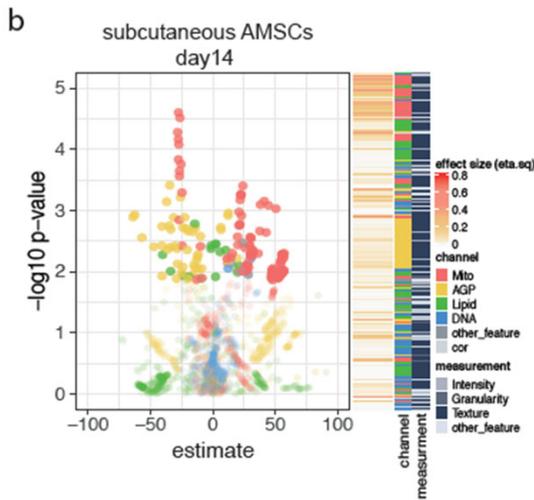
Mark Daly
 Hailiang Huang

Chuck Epstein



Connecting V2F for the 18q21 metabolic risk locus

Lipodystrophy-specific PRSs implicate mitochondrial hyper-activity and fission in subcutaneous adipocytes and lipid accumulation in visceral adipocytes



Laber S*, Strobel S*, et al manuscript in preparation

Accelerating Medicines Partnership

May 21, 2021

A heuristic approach to gene identification.....
Mark McCarthy & Anubha Mahajan, Oxford



43

Outline

- What are the things that we know that a model should include?
- How did we come up with the heuristic model, and what did we find?
- Are there any updates given subsequent data?
 - additional GWAS (DIAMANTE, MVP)
 - additional genomic data (islets, fat, muscle, liver...)
- How does the heuristic model compare to the other approaches?
- What should we do next?

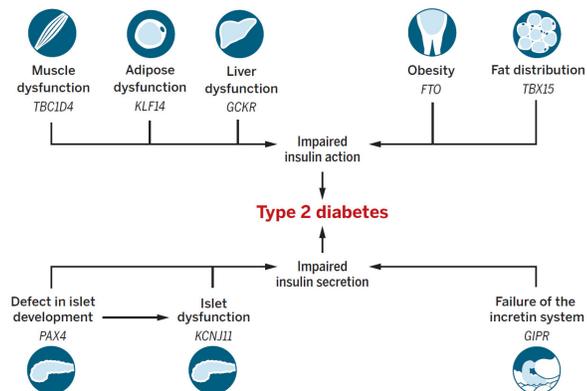


44

What do we know? (1)

Multiple cell types are involved, and multiple cell states (including developmental), mediate T2D risk through multiple processes

Common variants of modest effect can nudge T2D risk up or down by interfering with any one of several pathophysiological processes implicated in T2D risk



45

What do we know? (2)

Fine-mapping is getting ever better (better arrays, better reference panels, better methods, more transethnic data, more sequencing), and at many signals we can feel reasonably confident about which variant is causal.

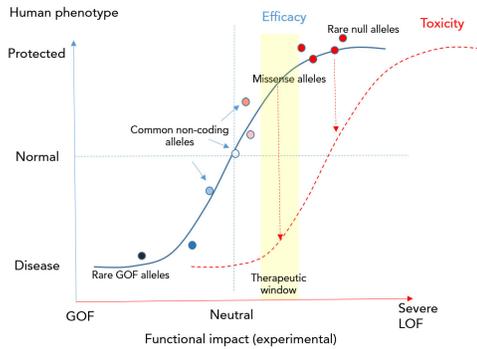
This really helps to locate the regulatory (or other) mechanisms through which signals operate, and (for regulatory variants) can home in on the cell-type, state and process involved

*but the reductionist approach will sometimes fail eg where it's a haplotype effect with multiple causal variants

46

What do we know? (3)

Around 10% of the overall signal resides in coding variants, but these are particularly informative from a mechanistic perspective, especially when the effect is large (which usually means the variant is rare)



But:

- Not every credible set coding variant is causal (Mahajan NG 2018a)
- The “equivalence” assumption may be violated: the phenotypic spectrum of coding & regulatory variants for the same gene may differ (due to tissue expression)
- The “proximity” assumption (ie that multiple signals at a locus operate thru the same mechanism) may be violated: as signal density increases, we become less confident that a regulatory and coding signal at the same “locus” have the same mechanism

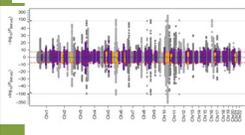


47

Causal coding variants

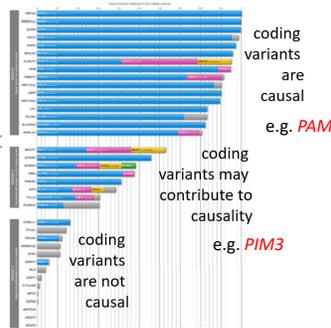
GWAS

Fine-mapping type 2 diabetes loci to single-variant resolution by using high-density imputation and islet-specific epigenome maps



e.g. *CDKN1B*
PATJ

Exome array



coding variants are causal e.g. *PAM*

coding variants may contribute to causality e.g. *PIM3*

coding variants are not causal

Exome sequencing

Single variant e.g. *PAX4*

Gene-based e.g. *SLC30A8*

Rare variant linkage

Monogenic e.g. *HNF1A*, *WFS1*

A mix of

- the main GWAS variant is a coding variant (*PAM*)
- the main GWAS signal is non-coding, but there's a secondary signal in the region that is coding (*NGN3*, *HNF1A*)
- there's an exome sequencing signal (single variant or burden) (*PAX4*)
- none of the GWAS signals is coding, but there's a rare coding variant in one of the genes nearby (*GCK*)

48

What do we know? (4)

Indeed, as density of signals gets ever higher, the concept of a “locus” is increasingly redundant

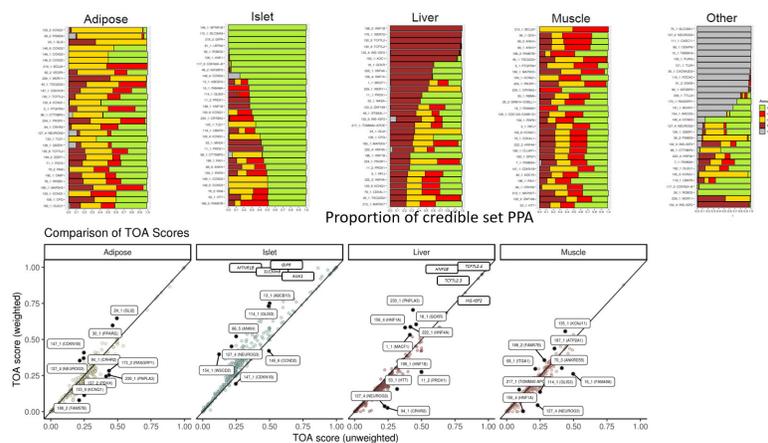
- Some loci involve 20+ signals (*INS/IGF2/KCNQ1*) – are all of these working through the same mechanism?
- Not all signals at the same locus influence the same cell type (*TCF7L2*)
- We have clear examples where there are two overlapping signals with distinct mechanisms (NKX6.3 via islets; ANK1 via insulin action tissues)
- Increasing evidence at some loci that more than one gene is involved (*STARD10 & ARAP1*)



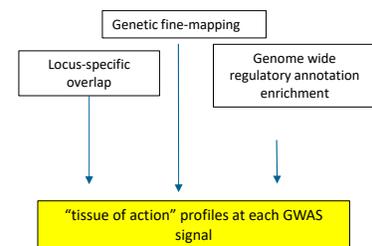
49

What do we know? (5)

Though the strongest signals for GWAS enrichment map to islet enhancers, that does not mean that all signals work through the islets: indeed getting to the right cell-type and state for each signal is critical for accurate functional inference



50



but:

- Incomplete fine-mapping
- Bulk tissue only
- Much more granular data needed

What do we know? (6)

There's a lot of **molecular pleiotropy**: the same non-codign variant may be associated with many different regulatory effects (different genes, different cell-types)

ARTICLE

A Multi-tissue Transcriptome Analysis of Human Metabolites Guides Interpretability of Associations Based on Multi-SNP Models for Gene Expression

Anne Ndungu,^{1,5} Anthony Payne,^{1,5} Jason M. Torres,^{1,5} Martijn van de Bunt,^{1,2,3,6} and Mark I. McCarthy^{1,2,4,6,*}

Used TWAS approach to "find", using GTEx data, the genes that we knew to be driving metabolomic QTL effects

- Sensitivity 67%
- Specificity 8% (↑ to 19% if co-localizing)
- Lots of "bystander" genes indistinguishable from the true gene

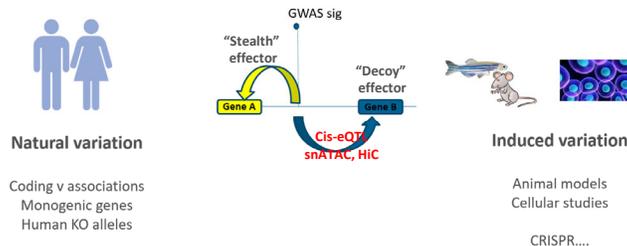
- Colocalization certainly helps to distinguish true from false (but is far from sufficient)
- Getting better cell-type specific data will help (both with functional fine-mapping, and then at finding the downstream effectors in the right cell type state).
- ABC model and other similar efforts to integrate data and model effects will certainly help
- Till then, caution required from using cis-eQTL data (and other V->G data) to assign effectors, esp if you don't know which cell-type you should be looking in (in general; and for that specific signal)



51

What do we know (7)

However confident we are about the connections between a GWAS causal variant and a downstream effector (based on some combination of cis-eQTL, snATAC, HiC, ABC...), these connections are correlative not causal. Sooner or later we need some perturbation data.



52

What do we know (8)

We have a wealth of data that can help us find causal variants, the regulatory elements they influence, the genes they regulate, and the consequences of perturbing variants, genes and pathways

but each of these comes with its own assumptions and uncertainties, and we need to build inference across all of these different data types

We ideally want to do that in a probabilistic way but we haven't yet worked out how to do that....

... Whilst we were waiting, a heuristic approach might help....



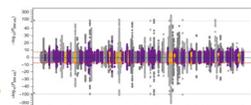
53

Sources of evidence available for a gene's classification

GENETIC EVIDENCE

- GWAS coding evidence
- Exome array evidence
- Burden test evidence
- Monogenic associations
- Other genetic evidence

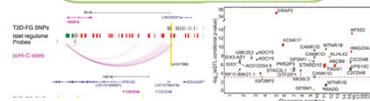
Mahajan et al Nature Genetics 2018a
Mahajan et al Nature Genetics 2018b
Flannick et al Nature 2019
OMIM



REGULATORY EVIDENCE

- Islet cis-eQTLs
- Other relevant cis-eQTLs
- Islet chromatin conformation
- Allelic imbalance
- Glucose regulation
- Other regulatory evidence

Van De Bunt et al PLoS Genetics 2015
Scott et al Nat Commun 2016
Ottosson-Laakso et al Diabetes 2017
Varshney et al PNAS 2017
Civelek et al AJHG 2017
Thurner et al eLife 2018
Greenwald et al Nat Commun 2019
Escalada et al Nat Commun 2019
Viñuela et al Nat Commun 2021



PERTURBATION EVIDENCE

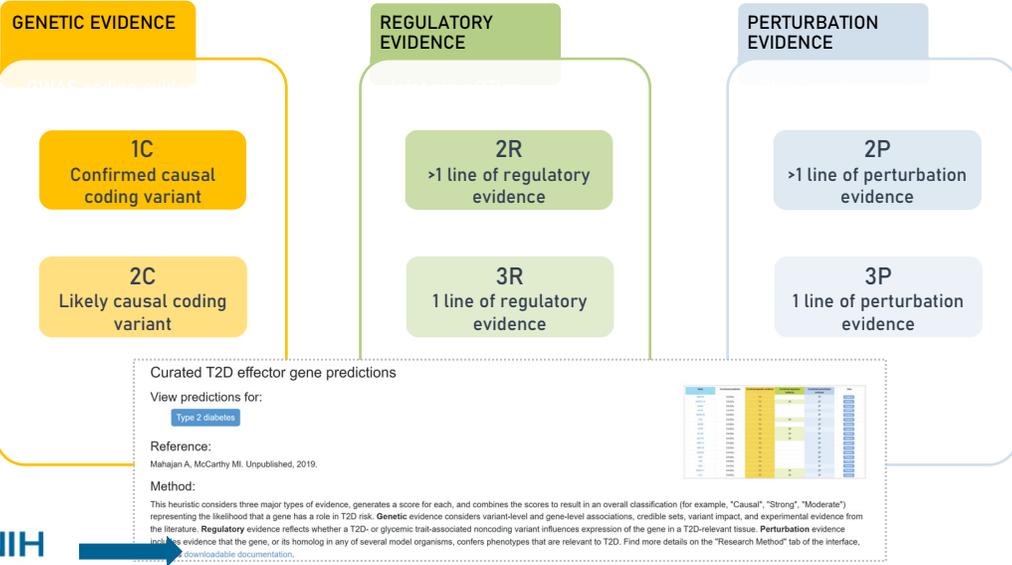
- RNA interference evidence
- Zebrafish mutant phenotype
- Mouse mutant phenotype
- Drosophila mutant phenotype
- Rat mutant phenotype
- Other perturbation evidence

Thomsen et al Diabetes 2016
Peiris et al Nat Commun 2018



54

Heuristic model



55

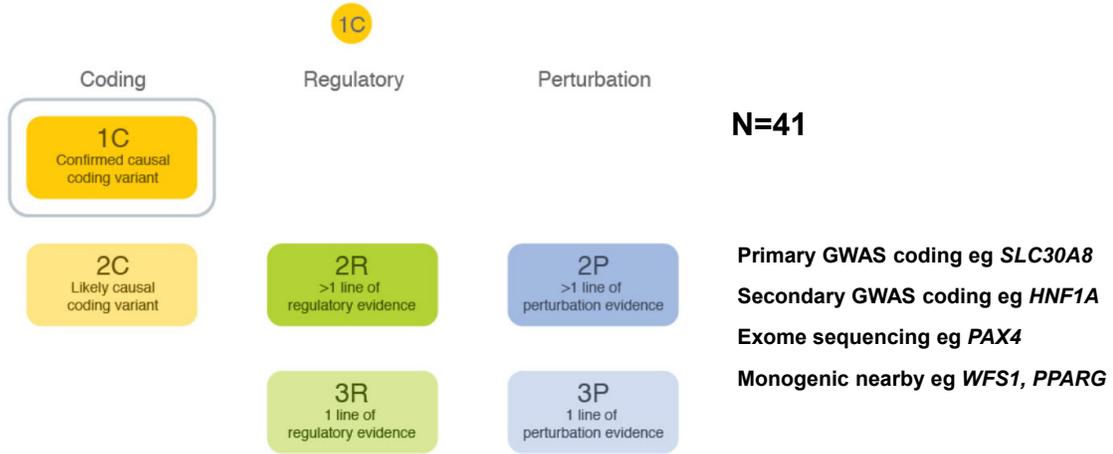
Evidence on the portal

Gene	Combined prediction	Combined genetic evidence	Combined regulatory evidence	Combined perturbation evidence	View
GLP1R	CAUSAL	1C	3R	2P	Evidence
Predicted T2D effector gene					
GLP1R					
GWAS coding evidence		Exome array evidence		Burden test evidence	Monogenic associations
Strong		Strong		Medium	PMID:27252175 PMID:29941447
Islet cis-eQTLs		Other relevant cis-eQTLs		Islet chromatin conformation	Allelic imbalance
RNA interference evidence		Zebrafish mutant phenotype		Mouse mutant phenotype	Drosophila mutant phenotype
		decreased body weight abnormal circulating insulin level abnormal glucose homeostasis abnormal pancreas secretion decreased circulating insulin level decreased lean body mass abnormal glucose tolerance impaired glucose tolerance abnormal food intake incre			Kat mutant phenotype
					decreased circulating glucose level increased heart rate increased systemic arterial blood pressure increased insulin secretion
Previously associated loci					
1					
Other regulatory evidence					
Other perturbation evidence					
2P					
Evidence					
Previously associated loci					
0					
Other regulatory evidence					
Other perturbation evidence					
PMID:24915262					
PMID:24915262					

Gene	Combined prediction	Combined genetic evidence	Combined regulatory evidence	Combined perturbation evidence	View
HNFI1A	CAUSAL	1C		2P	Evidence
Predicted T2D effector gene					
HNFI1A					
GWAS coding evidence		Exome array evidence		Burden test evidence	Monogenic associations
Strong		Strong		Strong	IMDY
Islet cis-eQTLs		Other relevant cis-eQTLs		Islet chromatin conformation	Allelic imbalance
RNA interference evidence		Zebrafish mutant phenotype		Mouse mutant phenotype	Drosophila mutant phenotype
increased insulin secretion		pancreatic B cell decreased amount pancreatic B cell decreased area + glucose homeostasis disrupted		abnormal liver morphology hepatocyte morphology decreased body weight decreased body size hyperglycemia increased urine glucose level hepatic steatosis decreased circulating insulin level increased circulating alanine transaminase level increased	Rat mutant phenotype
Other perturbation evidence					

56

Causal

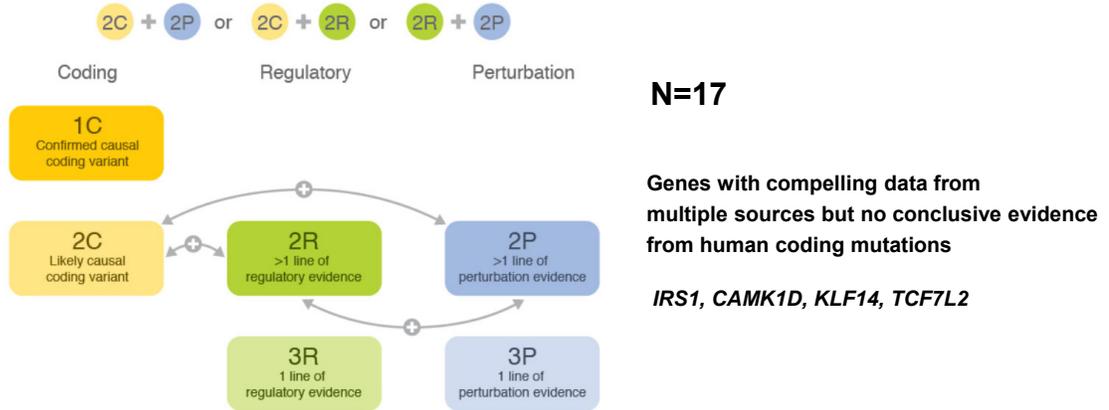


ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

57

Strong

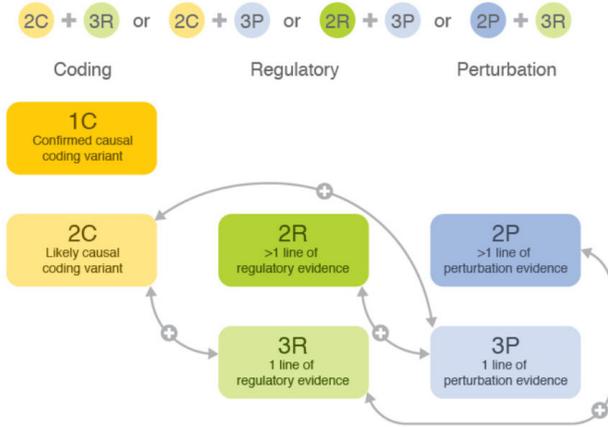


ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

58

Moderate



N=21

Genes with good data from multiple sources but no conclusive evidence from human coding mutations

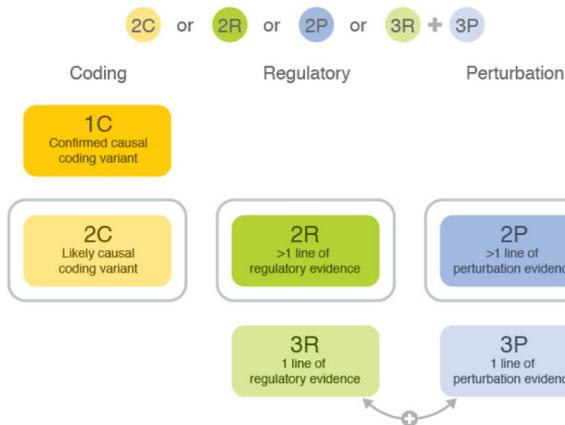
Eg *ADCY5*, *IGF2BP2*, *JAZF1*

ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

59

Possible



N=13

Genes that have more than 1 line of evidence but short of compelling

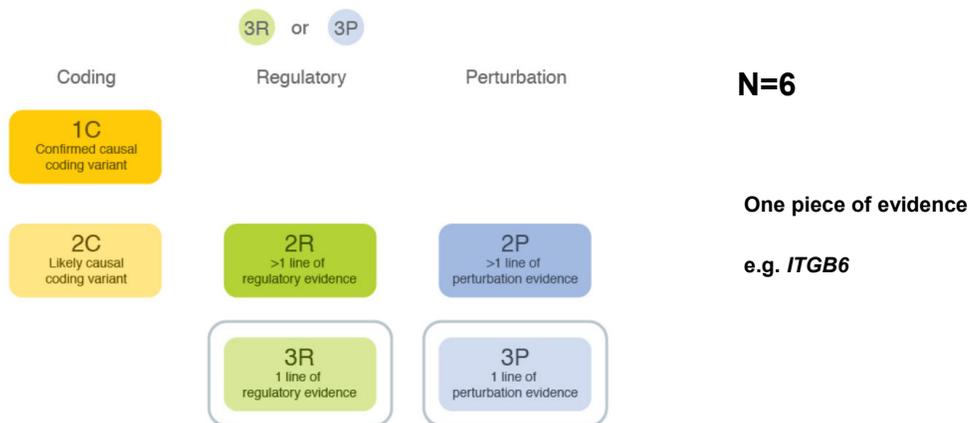
e.g. *ST6GAL1*

ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

60

Weak



ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

61

T2D-related

N=34

Not T2D loci, but implicated in similar processes by dint of either

- Likely causal for monogenic diabetes or related trait eg *CEL*, *AKT2*
- Evidence connecting to variation in continuous glycemic traits eg *GRB10*, *SIX2/3*

Potentially relevant to network construction, so good to have them on hand

ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

62

What has changed since then?

Over 600 regions in the genome influence an individual's lifetime risk for T2D

More T2D associated loci

August 2018

Zeggini et al 2008, *Nature Genetics*
 Morris et al 2012, *Nature Genetics*
 Gaulton et al 2015, *Nature Genetics*
 Scott et al 2017, *Diabetes*
 Mahajan et al 2018 (a), *Nature Genetics*
 Suzuki et al 2019, *Nature Genetics*
 Vujkovic et al 2020, *Nature Genetics*

Voight et al 2010, *Nature Genetics*
 Mahajan et al 2014, *Nature Genetics*
 Fuchsberger et al 2016, *Nature*
 Zhao et al 2017, *Nature Genetics*
 Mahajan et al 2018 (b), *Nature Genetics*
 Spracklen et al 2020, *Nature*
 Mahajan et al 2020, *medRxiv*

More regulatory data

Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D

Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density

Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS

Better fine-mapping resolution

Construct 99% "credible sets" of variants most likely to be causal

Log10 number of variants in 99% credible set

Log10 length of 99% credible set (kb)

241-OR: Causal Gene Candidates for Type 2 Diabetes Based on Protein-Coding Variants in 127,676 Individuals

PETER DORNBOG, LAURA RAFFIELD, XIANYONG YIN and JASON FLANNICK

Tissue of action

Gene	Islet	Adipose	Liver	Unclassified
ANKK1	High	Low	Low	Low
ANKK2	High	Low	Low	Low
ANKK3	High	Low	Low	Low
ANKK4	High	Low	Low	Low
ANKK5	High	Low	Low	Low
ANKK6	High	Low	Low	Low
ANKK7	High	Low	Low	Low
ANKK8	High	Low	Low	Low
ANKK9	High	Low	Low	Low
ANKK10	High	Low	Low	Low
ANKK11	High	Low	Low	Low
ANKK12	High	Low	Low	Low
ANKK13	High	Low	Low	Low
ANKK14	High	Low	Low	Low
ANKK15	High	Low	Low	Low
ANKK16	High	Low	Low	Low
ANKK17	High	Low	Low	Low
ANKK18	High	Low	Low	Low
ANKK19	High	Low	Low	Low
ANKK20	High	Low	Low	Low
ANKK21	High	Low	Low	Low
ANKK22	High	Low	Low	Low
ANKK23	High	Low	Low	Low
ANKK24	High	Low	Low	Low
ANKK25	High	Low	Low	Low
ANKK26	High	Low	Low	Low
ANKK27	High	Low	Low	Low
ANKK28	High	Low	Low	Low
ANKK29	High	Low	Low	Low
ANKK30	High	Low	Low	Low
ANKK31	High	Low	Low	Low
ANKK32	High	Low	Low	Low
ANKK33	High	Low	Low	Low
ANKK34	High	Low	Low	Low
ANKK35	High	Low	Low	Low
ANKK36	High	Low	Low	Low
ANKK37	High	Low	Low	Low
ANKK38	High	Low	Low	Low
ANKK39	High	Low	Low	Low
ANKK40	High	Low	Low	Low
ANKK41	High	Low	Low	Low
ANKK42	High	Low	Low	Low
ANKK43	High	Low	Low	Low
ANKK44	High	Low	Low	Low
ANKK45	High	Low	Low	Low
ANKK46	High	Low	Low	Low
ANKK47	High	Low	Low	Low
ANKK48	High	Low	Low	Low
ANKK49	High	Low	Low	Low
ANKK50	High	Low	Low	Low
ANKK51	High	Low	Low	Low
ANKK52	High	Low	Low	Low
ANKK53	High	Low	Low	Low
ANKK54	High	Low	Low	Low
ANKK55	High	Low	Low	Low
ANKK56	High	Low	Low	Low
ANKK57	High	Low	Low	Low
ANKK58	High	Low	Low	Low
ANKK59	High	Low	Low	Low
ANKK60	High	Low	Low	Low
ANKK61	High	Low	Low	Low
ANKK62	High	Low	Low	Low
ANKK63	High	Low	Low	Low
ANKK64	High	Low	Low	Low
ANKK65	High	Low	Low	Low
ANKK66	High	Low	Low	Low
ANKK67	High	Low	Low	Low
ANKK68	High	Low	Low	Low
ANKK69	High	Low	Low	Low
ANKK70	High	Low	Low	Low
ANKK71	High	Low	Low	Low
ANKK72	High	Low	Low	Low
ANKK73	High	Low	Low	Low
ANKK74	High	Low	Low	Low
ANKK75	High	Low	Low	Low
ANKK76	High	Low	Low	Low
ANKK77	High	Low	Low	Low
ANKK78	High	Low	Low	Low
ANKK79	High	Low	Low	Low
ANKK80	High	Low	Low	Low
ANKK81	High	Low	Low	Low
ANKK82	High	Low	Low	Low
ANKK83	High	Low	Low	Low
ANKK84	High	Low	Low	Low
ANKK85	High	Low	Low	Low
ANKK86	High	Low	Low	Low
ANKK87	High	Low	Low	Low
ANKK88	High	Low	Low	Low
ANKK89	High	Low	Low	Low
ANKK90	High	Low	Low	Low
ANKK91	High	Low	Low	Low
ANKK92	High	Low	Low	Low
ANKK93	High	Low	Low	Low
ANKK94	High	Low	Low	Low
ANKK95	High	Low	Low	Low
ANKK96	High	Low	Low	Low
ANKK97	High	Low	Low	Low
ANKK98	High	Low	Low	Low
ANKK99	High	Low	Low	Low
ANKK100	High	Low	Low	Low

ABC model

Precision (PPV, %)

Recall (sensitivity, %)

● All genes within 100 kb
 ● All genes within 100 kb
 ● Other predictors

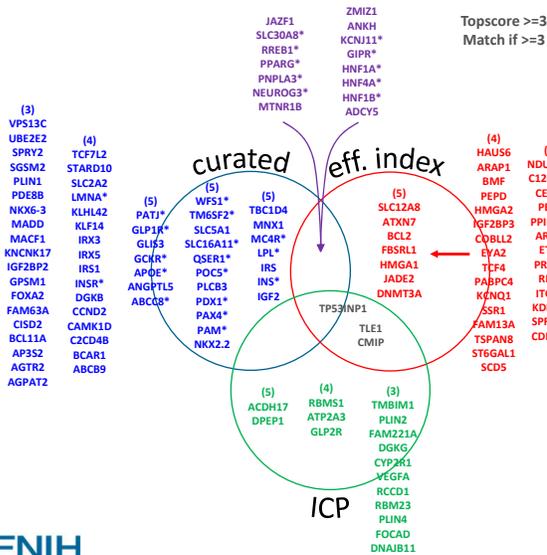
Expanded candidate effector transcript list

Coding:	19 signals including <i>MYO5C</i> & <i>ACVR1C</i>					
Protein QTLs:	PAM, PGM1, PRSS3, APOE (7 signals at 7 regions)					
RNA QTLs	Islets (42)	SAT (28)	VAT (18)	SM (23)	Liver 8	Brain 3
126 cis-eqtl at 66 signals	<i>CEP68</i> <i>STARD10</i> <i>DKKB</i> <i>NKX6.3</i> <i>ST6GAL1</i>	<i>CEP68</i> <i>IRS1</i> <i>KLF14</i> <i>ANK1</i> <i>PLEKH1</i> <i>PRC1</i>	<i>CEP68</i> <i>IRS1</i> <i>INHBB</i> <i>PCGF3</i> <i>JAZF1</i>	<i>CEP68</i> <i>ANK1</i> <i>ZZEF1</i> <i>PCGF3</i> <i>JAZF1</i> <i>SKOR1</i>	<i>CEP68</i> <i>JAZF1</i> <i>SLC22A3</i>	<i>CLAU1</i>
Chromatin Interactions	Islets		SAT		Liver	
214 contacts	<i>TCF7L2</i> <i>RFX3</i> <i>ST6GAL1</i> <i>CTRB2</i> <i>KCNK17</i>		<i>CCND2</i> <i>IRX5</i> <i>GIPR</i> <i>PPARG</i> <i>TCF7L2</i> <i>ANK1</i>		<i>CMIP</i> <i>CCND2</i> <i>PPARG</i> <i>SLC22A3</i> <i>FOXA2</i>	

Curation of the next version of type 2 diabetes effector gene predictions using the heuristic approach underway.....



Comparisons from the portal



	curated	EI	ICP
1: 40Mb	MACF1	PABPC4	
2: 65Mb		CEP68 SPRED2	
3: 185Mb	IGF2BP2	STGGAL1	DGKG DNAJB11
5: 102Mb	PAM	PPIPSK2	
6: 7Mb	RREB1	RREB1 SSR	
6: 39Mb	GLP1R KCNK17		
7:23Mb		IGF2BP3	FAM221A
8: 95Mb		TP53INP1 NDUFAF6	TP53INP1
9:19Mb		HAU56	PLIN2
9:139Mb	GPSM1 AGPAT2		
11:2Mb	IGF2 INS	KCNQ1	
11:72Mb	STARD10	ARAP1	
12: 121Mb	HNF1A	HNF1A KDM2B	
15: 62Mb	VPS13C C2CD4B		
15: 90Mb	PLIN1 AP3S2		
16: 54 Mb	IRX3, IRX5		
19: 46 Mb	APOE GIPR	GIPR	



Some reasons for discrepant assignments..

- coding variants only account for 10% of overall GWAS signal, but feature disproportionately amongst “solved” loci (in part because we borrow information from monogenic genes) and so currently provide a lot of the evidence base;
- methods based on regulatory V→G analyses alone miss these, including several “slam dunk” signals such as *LMNA*, *PAM*, *MC4R*, *TM6SF2* etc;
- EI picks up several coding variant signals (eg *PPARG*, *SLC30A8*, *HNFs*) but some reflect loci where there is both a coding variant signal, AND a regulatory signal (*PPARG* is a good example): EI is picking up the latter;
- EI (per the portal) used the BMI-adjusted T2D signal so misses some loci like *MC4R* and *FTO* mediated through BMI; also the portal only reports the top 100 predictions for EI in the combined evidence, which underestimates support (and suggests discordance) for some genes that EI did in fact support;
- ICP (as we understand it) excluded signals w coding variants in the credible set (as does ABC model) so removes many of the strongest signals – this isn’t clear on the portal, and leaves many slam dunk signals looking bereft of support;
- several regulatory signals where (we feel) the evidence is extremely strong – even in the absence of coding variants -- are not detected by ICP or EI - examples include *KLF14* & *DGKB*: this is worth investigating;
- many of the apparent discrepancies on the table on the previous slide have prosaic explanations: eg at some, the different methods focused on different GWAS signals in the same region (the entries in the table were only mapped based on gene locations)



Limitations of our approach

- arbitrary & heuristic
- limited “ground truth” against which to compare
 - (NB: monogenic genes already baked in)
- directional consistency not fully taken into account (eg around animal models)
- hard to find or adjudicate all literature: subjective decisions about which published material is worthy of inclusion
- new data arriving all the time (not a static situation); as not automated can be hard to update
- we don't know what pieces of data should carry more weight than others
- we may end up overfittig to existing understanding (if existing understanding is incorrect)

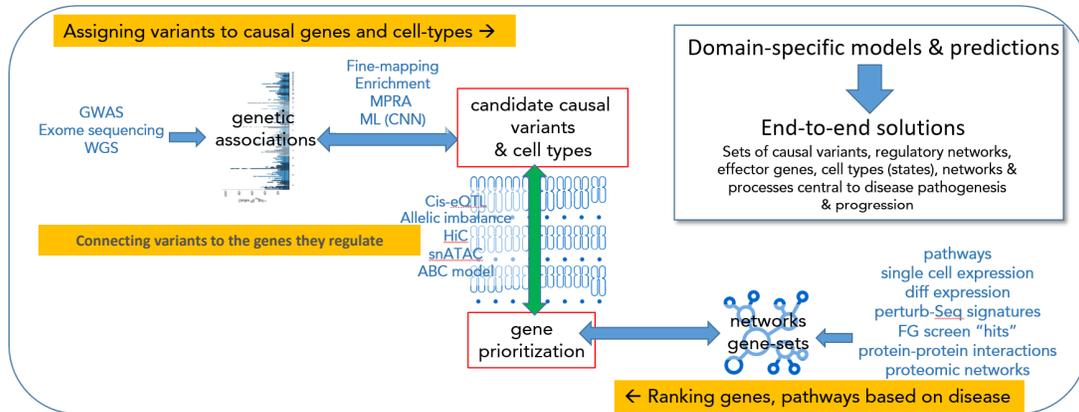
But: it seems to be consistent with expert consensus at most signals, and seems to include more relevant data than methods that are focused on regulatory variants alone



Can we do better?

- ❑ **More, better data about each gene:**
 - ❑ regulation: single-cell data; snATAC;
 - ❑ perturbation: exomes/genomes; null-alleles; screens
 - ❑ multivariate phenotypic readouts: PheWAS, proteomic signatures etc....
- ❑ **Data type “bake-offs”**
 - downweight or discard divergent data types
- ❑ **Scoring systems** for combining evidence
 - eg **Open Targets** (harmonic sum) $S_{L,i} = S_1 + \frac{S_2}{2^2} + \frac{S_3}{3^2} + \frac{S_4}{4^2} \dots + \frac{S_i}{i^2}$ **Allelic series: dose-response**
- ❑ **Truth sets** – but none is perfect
 - * monogenic genes; drug targets; transfer learning from other diseases/traits?
- ❑ Integrate information on “emergent” properties: eg **connectivity in PPI** or pathway space, or genome-wide screens
- ❑ **Machine learning:** supervised; knowledge graphs

Next steps



Weeks et al – POPS
 Jagadeesh et al – sc and GWAS
 Fernandez-Tajes et al – GWAS & PPI data



Partners for Innovation, Discovery, Health | www.fnih.org

69

69

Acknowledgements



- Andrew Morris
- Ines Barroso
- Xueling Sim & the E Asian team
- John Chambers & the S Asian team
- Maggie Ng & MEDIA
- Ayesha Motala, Manj Sandhu
- Piper Below & Hispanic T2D team
- Jerry Rotter
- James Meigs & the TOPMED team
- Jose Florez
- Mike Boehnke
- Everyone else in DIAGRAM, DIAMANTE, MAGIC, GIANT

- Anna Gloyn
- Steve Parker
- Karen Mohlke
- Kyle Gaulton
- Rob Sladek
- Kasper Lage
- Melina Claussnitzer
- Maïke Sander
- Bing Ren
- Patrick MacDonald
- Jesse Engreitz
- Jason Flannick
- Manolis Dermitzakis
- Eric Fauman
- Robert Plenge
- Sally John



ACCELERATING MEDICINES PARTNERSHIP (AMP)
 TYPE 2 DIABETES

70



Visualization of target prioritization tools on the AMP Common Metabolic Diseases Knowledge Portal

Noël Burt & Jason Flannick
May 21, 2021



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

72

Today



- Representation of effector prediction approaches
- Enhancing & expanding effector prediction approaches
- Decision support & advancing tools for visualization for gene & target prioritization
- Future directions



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

73

Evolution of the portal

Dataset-level genetic association results

ACCELERATING MEDICINES PARTNERSHIP (AMP)
 TYPE 2 DIABETES GENETICS beta

HOME · ABOUT THE DATA · TUTORIAL · POLICIES · CONTACT · FORUM MANAGE · MARYC@BROADINSTITUTE.ORG · LOG OUT

TCF7L2

Uniprot Summary: Participates in the Wnt signaling pathway and modulates MYC expression by binding to its promoter in a sequence-specific manner. Acts as repressor in the absence of CTNNB1, and as activator in its presence. Activates transcription from promoters with several copies of the Tcf motif 5'-CCTTTGATC-3' in the presence of CTNNB1. TLE1, TLE2, TLE3 and TLE4 repress transactivation mediated by TCF7L2/TCF4 and CTNNB1. Expression of dominant-negative mutants results in cell-cycle arrest in G1. Necessary for the maintenance of the epithelial stem-cell compartment of the small intestine.

➤ Variants and associations

Explore variants within 100kb of TCF7L2

Click on a number below to generate a table of variants associated with type 2 diabetes in the following categories:

data type	sample size	total variants	genome-wide significant variants P < 5e-8	locus-wide significant variants P < 1e-5	nominal significant variants P < 0.05
GWAS	69,033	230	32	58	123
exome chip	79,854	3	0	0	0
exome sequence	16,760	161	0	0	5

Variants within 100kb of TCF7L2 are also genome-wide significantly associated with:

- two-hour glucose
- fasting glucose
- fasting proinsulin

➤ Explore significant variants with IGV

2015



Evolution of the portal

Distilled gene-level results

MC4R Gene Page guide

Strong evidence for signal Look for another gene Go

Uniprot Summary: Receptor specific to the heptapeptide core common to adrenocorticotrophic hormone and alpha-, beta-, and gamma-MSH. This receptor is mediated by G proteins that stimulate adenylate cyclase.

PHENOTYPES WITH SIGNALS

BMI Height HDL cholesterol Type 2 diabetes Coronary artery disease Triglycerides Waist circumference
 Subcutaneous adipose tissue volume Creatinine Fasting glucose &HbA1c eGFR-creat (serum creatinine) Waist-hip ratio
 Visceral adipose tissue attenuation Cholesterol Urinary albumin-to-creatinine ratio Fasting glucose > Additional phenotypes...

Note: traits from the Oxford Biobank exome chip dataset are not currently included in this analysis.

MC4R is located on chromosome 18 between position 57938514 and 58140051

Common variants: Type 2 diabetes High-impact variants: Type 2 diabetes Credible sets: Type 2 diabetes

The Common variants tab shows information about variants associated with the selected phenotype whose minor allele frequency (MAF) is greater than 5%.

Variant ID	dbSNP ID	Major allele	Minor allele	Predicted impact	p-Value	Effect	MAF	Data set
18:57864720.G.A	rs12970134	G	A	Intergenic_variant	1.19e-8	1.08	1.0	DIAGRAM GWAS + MetaboChip
18:57858829.T.C	rs8089364	T	C	upstream_gene_variant	1.53e-8	1.08	1.0	DIAGRAM GWAS + MetaboChip
18:57859543.C.A	rs12969709	C	A	upstream_gene_variant	1.56e-8	1.08	1.0	DIAGRAM GWAS + MetaboChip
18:57859749.C.A	rs171312	C	A	Intergenic_variant	2.10e-8	1.08	1.0	DIAGRAM Transethnic meta-analysis
18:57811330.A.G	rs12966550	A	G	Intergenic_variant	2.50e-8	1.08	1.0	DIAGRAM Transethnic meta-analysis
18:57849023.G.A	rs12967135	G	A	Intergenic_variant	2.80e-8	1.09	1.0	DIAGRAM Transethnic meta-analysis
18:57812989.A.G	rs12955983	A	G	Intergenic_variant	3.20e-8	1.08	1.0	DIAGRAM Transethnic meta-analysis
18:57861943.G.A	rs1457489	G	A	upstream_gene_variant	3.48e-8	1.08	1.0	DIAGRAM GWAS + MetaboChip
18:57859618.C.G	rs8083289	C	G	Intergenic_variant	3.50e-8	1.08	1.0	DIAGRAM Transethnic meta-analysis
18:57851097.T.C	rs28680381	T	C	Intergenic_variant	3.60e-8	1.08	1.0	DIAGRAM Transethnic meta-analysis

2017



Evolution of the portal

Top variants: Type 2 diabetes | High-impact variants: Type 2 diabetes | Credible sets: Type 2 diabetes | Genes in region: Type 2 diabetes

This tab displays annotations and the results of computational methods that integrate genetic association results, genetic linkage, tissue-specific expression data and eQTLs, co-expression data, and other biological annotations for this genomic region. This information can help researchers to decide which of the genes surrounding a genetic association signal is most likely to account for that signal, suggesting its potential involvement in a trait or disease of interest. Click this question mark to see tips on navigating the table; click the question marks next to each method or annotation type in the table below to see a brief description; or download complete documentation. Please note that although the methods have been run according to published best practices, the results shown here are experimental.

Results of methods to support gene prioritization

		CCDC3 ID: 12938577-13141702	CAMK1D ID: 12291431-12877595	CDC123 ID: 12237915-12292637
Significance of association	Firth gene associations	p=0.987 (Extreme p-value aggregation test)	p=0.822 (Extreme p-value aggregation test)	p=0.884 (Extreme p-value aggregation test)
Significance of association	SKAT gene associations	p=0.541 (Extreme p-value aggregation test)	p=0.973 (Extreme p-value aggregation test)	p=0.689 (Extreme p-value aggregation test)
Significance of association	MetaXcan	p=0.0932	p=3.35e-31	p=0.00000272
Significance of association	DEPICT gene sets		p=0.0000263 (GO:0004713)	
Significance of association	DEPICT gene prioritization			p=0.0322
Posterior probability	eCAVIAR		CLPP=1.00 (Lower leg skin sun exposed)	CLPP=1.00 (Tibial artery)
Posterior probability	COLOC		CLPP=1.00 (Lung)	CLPP=0.253 (Lower leg skin sun exposed)
Annotation	Mouse knockout phenotypes			records=4
Annotation	T2D effector gene list		STRONG	

2019



Converting to Knowledge

Building Effector Gene Knowledge



Tools for Prioritization



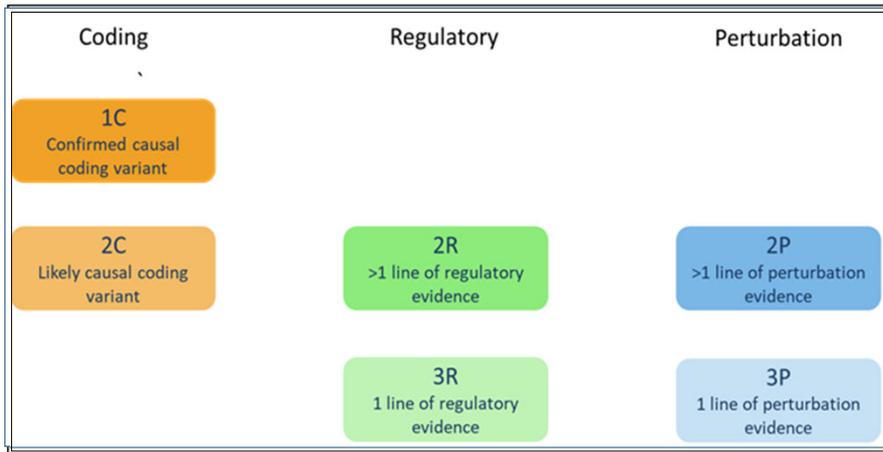
Curated Effector Transcript Heuristic



Anubha Mahajan



Mark McCarthy



1 Coding Evidence

2 Regulatory Evidence

3 Perturbation Evidence

ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

78

Where to find the results?

ACCELERATING MEDICINES PARTNERSHIP (AMP)

CMDKP

Home Data Tools KP Labs Information Contact Login

Gene Finder
Predicted Effector Genes
T2D Effector Prediction Summary

COMMON METABOLIC DISEASES KNOWLEDGE PORTAL

Providing data and understanding and treatment of common metabolic diseases

Gene, region or variant Phenotypes Disease-specific portals

Search

examples: PCSK9, rs1260326, chr9:21,940,000-22,190,000

79

Summary & Documentation

Contributing research methods

Curated T2D effector gene predictions

View predictions for:

[Type 2 diabetes](#)

Reference:

Mahajan A, McCarthy MI. Unpublished, 2019.

Method:

This heuristic considers three major types of evidence, generates a score for each, and combines the scores to result in an overall classification (for example, "Causal", "Strong", "Moderate") representing the likelihood that a gene has a role in T2D risk. **Genetic** evidence considers variant-level and gene-level associations, credible sets, variant impact, and experimental evidence from the literature. **Regulatory** evidence reflects whether a T2D- or glycemic trait-associated noncoding variant influences expression of the gene in a T2D-relevant tissue. **Perturbation** evidence includes evidence that the gene, or its homolog in any of several model organisms, confers phenotypes that are relevant to T2D. Find more details on the "Research Method" tab of the interface, or in this [downloadable documentation](#).

Gene	Genetic evidence	Regulatory evidence	Perturbational evidence	Overall prediction
ANKRD1A	Yes	Yes	Yes	Causal
ANKRD1A	Yes	Yes	No	Strong
ANKRD1A	Yes	No	Yes	Strong
ANKRD1A	Yes	No	No	Moderate
ANKRD1A	No	Yes	Yes	Strong
ANKRD1A	No	Yes	No	Moderate
ANKRD1A	No	No	Yes	Moderate
ANKRD1A	No	No	No	None
ANKRD1A	Yes	Yes	Yes	Causal
ANKRD1A	Yes	Yes	No	Strong
ANKRD1A	Yes	No	Yes	Strong
ANKRD1A	Yes	No	No	Moderate
ANKRD1A	No	Yes	Yes	Strong
ANKRD1A	No	Yes	No	Moderate
ANKRD1A	No	No	Yes	Moderate
ANKRD1A	No	No	No	None
ANKRD1A	Yes	Yes	Yes	Causal
ANKRD1A	Yes	Yes	No	Strong
ANKRD1A	Yes	No	Yes	Strong
ANKRD1A	Yes	No	No	Moderate
ANKRD1A	No	Yes	Yes	Strong
ANKRD1A	No	Yes	No	Moderate
ANKRD1A	No	No	Yes	Moderate
ANKRD1A	No	No	No	None

80

Visualize the model & documentation

Curated T2D effector gene predictions

Phenotype
Type 2 diabetes

[View data](#) [View research method](#)

Gene

Prediction

Genetic evidence

Regulatory evidence

Perturbational evidence

Causal

Coding

1C
Confirmed causal coding variant

2C
Likely causal coding variant

Regulatory

3R
1 line of regulatory evidence

2R
>1 line of regulatory evidence

Perturbation

3P
1 line of perturbation evidence

2P
>1 line of perturbation evidence

81

Visualize & query the results

[View data](#) [View research method](#)

Gene	Prediction	Genetic evidence	Regulatory evidence	Perturbational evidence
<input type="text"/>				

*Click 'Evidence' button to view evidence data. *Hover evidence tables to see evidence group names.

Show all feature rows
 Hide feature headers
 Hide top level rows

Gene	Combined prediction	Combined genetic evidence	Combined regulatory evidence	Combined perturbation evidence	View
ABCB9	STRONG	2C	2R	3P	Evidence
ABCC8	CAUSAL	1C		2P	Evidence
ABO	WEAK		3R		Evidence
ADCY5	MODERATE		2R	3P	Evidence
ADRA2A	(T2D_related)		2R	2P	Evidence
ADRA2B	MODERATE	2C		2P	Evidence

82

Visualize the evidence

Gene	Combined prediction	Combined genetic evidence	Combined regulatory evidence	Combined perturbation evidence	View
NEUROG3	CAUSAL	1C		3P	Evidence
NKX2-2	CAUSAL	1C		2P	Evidence
NKX6-3	MODERATE		2R	3P	Evidence
PAM	CAUSAL	1C	2R	2P	Evidence
Predicted T2D effector gene			Previously associated loci		
PAM					
GWAS coding evidence	Exome array evidence	Burden test evidence	Monogenetic associations	Other genetic evidence	
Strong	Strong	Strong		PMID:24464100	
Islet cis-eQTLs	Other relevant cis-eQTLs	Islet chromatin conformation	Allelic imbalance	Glucose regulation	Other regulatory evidence
				1	PMID:30054598
RNA interference evidence	Zebrafish mutant phenotype	Mouse mutant phenotype	Drosophila mutant phenotype	Rat mutant phenotype	Other perturbation evidence
	whole organism decreased life span ciliary basal body-plasma membrane docking disrupted peptidylglycine monoxygenase activity disrupted	impaired glucose tolerance increased total body fat amount			PMID:30054598

83

Effector Index Predictions



Brent Richards

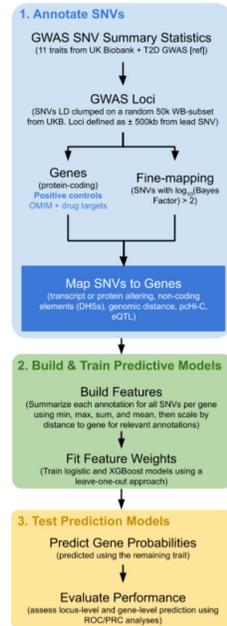


Vince Forgetta

An Effector Index to Predict Causal Genes at GWAS Loci

Vincenzo Forgetta¹, Lai Jiang^{1,8}, Nicholas A. Vulpesu², Megan S. Hogan², Siyuan Chen^{1,8}, John A. Morris^{1,3,4,10}, Stepan Grinek², Christian Benner², Mark I McCarthy², Eric Fauman⁷, Cecilia MT Greenwood^{1,8,9,10}, Matthew T. Maurano^{1,2}, J. Brent Richards^{1,8,10,11,12}

- Effector Index (Ei) algorithm calculates the probability of causality for each gene at a locus that harbors genome-wide significant SNVs for a disease or trait.
- Similar evidence sources for T2D, & 11 additional traits, *pre-print access*

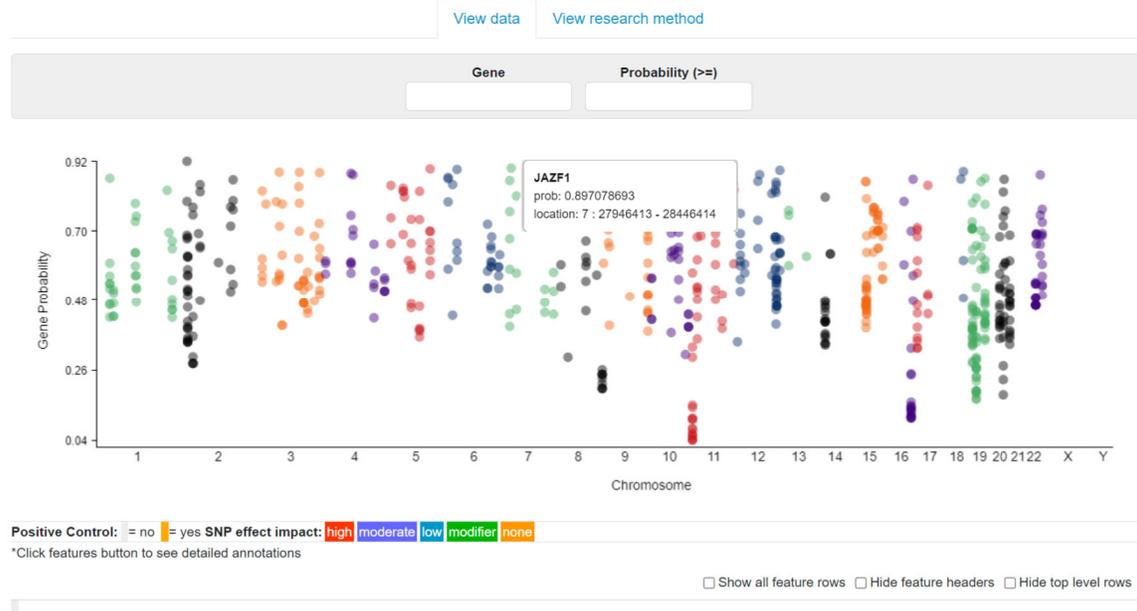


ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

84

Visualize & query the results



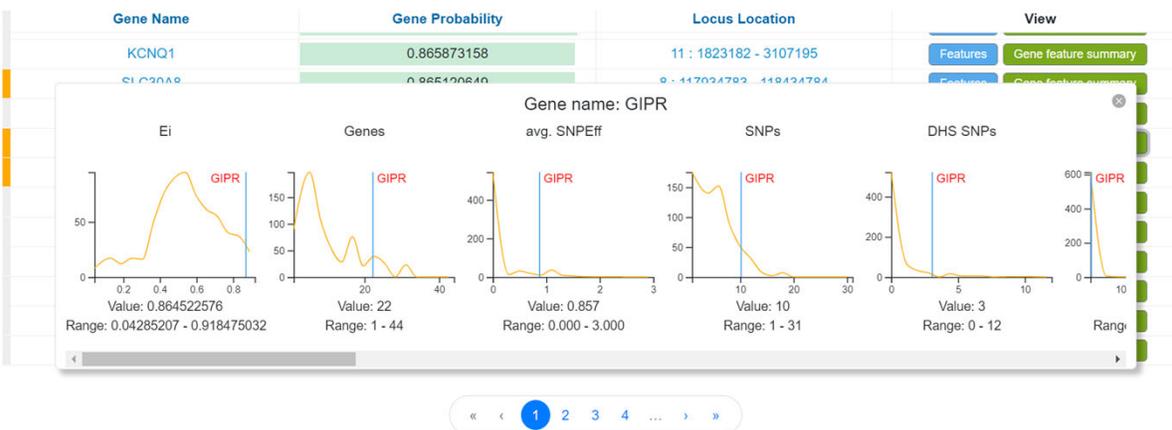
85

Visualize the evidence

Gene Name	Gene Probability	Locus Location	View									
TCF4	0.862573981	18 : 52837984 - 53337985	Features	Gene feature summary								
CMIP	0.862145782	16 : 81284790 - 81784791	Features	Gene feature summary								
EYA2	0.861776114	20 : 45348564 - 45848565	Features	Gene feature summary								
TP53INP1	0.860884726	8 : 95435147 - 96210768	Features	Gene feature summary								
<table border="1"> <thead> <tr> <th>gene.gid</th> <th>gene.eid</th> <th>gene.strand</th> <th>gene.tss</th> </tr> </thead> <tbody> <tr> <td>94241</td> <td>ENSG00000164938</td> <td>-</td> <td>95961639</td> </tr> </tbody> </table>					gene.gid	gene.eid	gene.strand	gene.tss	94241	ENSG00000164938	-	95961639
gene.gid	gene.eid	gene.strand	gene.tss									
94241	ENSG00000164938	-	95961639									
Ei	Genes	avg. SNPEff	SNPs	DHS SNPs	SNP dist.	Locus z-score	log10(BF)	GeneLen	PostPr/dist	Gene SNPEff	Gene z score	
0.860884726	9	1.000	8	4	339	6.892	6.401	23439	0.0000976	1.000	6.892	
snp name	snp locus	snp pos	maf	beta	se	z	prob	log10bf	log10bf group	snpEff impact	dbSNP fu	
rs1320164	L108	95960766	0.5	0.051	0.007	6.892	0.110	2.132	3.938	MODIFIER	intron,ne gene-	
rs7003387	L108	95961978	0.5	0.051	0.007	6.892	0.110	2.135	3.938	MODIFIER	none	
rs896852	L108	95960885	0.5	0.051	0.007	6.892	0.110	2.132	3.938	MODIFIER	intron,ne gene-	
rs896854	L108	95960510	0.5	0.051	0.007	6.892	0.111	2.137	3.938	MODIFIER	intron,ne gene-	
<table border="1"> <thead> <tr> <th>gene.gid</th> <th>gene.eid</th> <th>gene.strand</th> <th>gene.tss</th> </tr> </thead> <tbody> <tr> <td>94241</td> <td>ENSG00000164938</td> <td>-</td> <td>95961639</td> </tr> </tbody> </table>					gene.gid	gene.eid	gene.strand	gene.tss	94241	ENSG00000164938	-	95961639
gene.gid	gene.eid	gene.strand	gene.tss									
94241	ENSG00000164938	-	95961639									
COBL1	0.860340297	2 : 165258389 - 166012010	Features	Gene feature summary								

86

Visualize the evidence



87

Effector Index predictions

For each phenotype, click on "Top 100" to view the top 100 gene predictions (for fastest page loading) or on "Full list" to view all predictions.



Cardiovascular and lipid phenotypes

- Diastolic blood pressure
- LDL cholesterol
- Systolic blood pressure
- Triglycerides

Glycemic phenotypes

- Type 2 diabetes
- Random glucose

Musculoskeletal phenotypes

- Calcium
- eBMD (estimated bone mineral density)

Other phenotypes

- Bilirubin
- Hypothyroidism
- Red blood cell count

88

Integrated Classifier Predictions

- Integrated classifier predicts which genes are likely relevant to T2D risk within T2D GWAS loci through functional & semantic data
- Integrates ABC predictions
- Results made accessible even before *bioRxiv* submission



Alisa Manning



Tim Majarian



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

89

Visualize & query the results

[View data](#) [View research method](#)

Gene Locus id Positive probability

*Click 'Features' button to view feature data

Show all feature rows
 Hide feature headers
 Hide top level rows

Gene name	Locus id	Positive probability	View	
DPEP1	SPG7_16_89564055	0.84	Features	Gene features summary
PCDH17	PCDH17_13_58965435	0.8	Features	Gene features summary
TP53INP1	TP53INP1_8_95685147	0.8	Features	Gene features summary
TP53INP1	TP53INP1_8_96092422	0.8	Features	Gene features summary
PCDH17	SRGAP2D_13_59077406	0.76	Features	Gene features summary
GLP2R	GLP2R_17_9785187	0.72	Features	Gene features summary
TP53INP1	TP53INP1_8_95961626	0.72	Features	Gene features summary
ATP2A3	77FF1_17_3860356	0.68	Features	Gene features summary

90

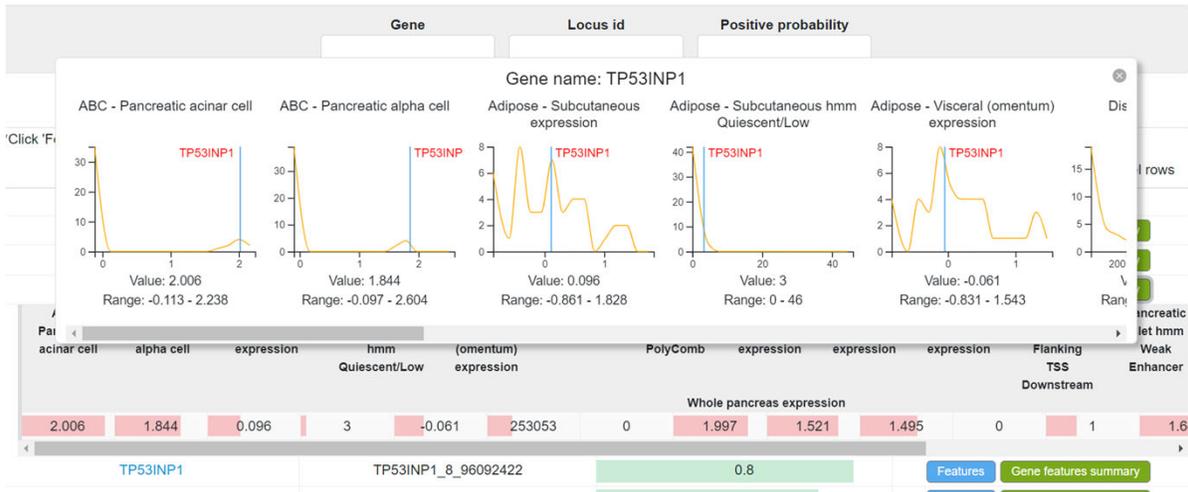
Visualize the evidence

Show all feature rows
 Hide feature headers
 Hide top level rows

Gene name	Locus id	Positive probability	View									
DPEP1	SPG7_16_89564055	0.84	Features	Gene features summary								
PCDH17	PCDH17_13_58965435	0.8	Features	Gene features summary								
TP53INP1	TP53INP1_8_95685147	0.8	Features	Gene features summary								
Whole pancreas expression												
ABC - Pancreatic acinar cell	ABC - Pancreatic alpha cell	Adipose - Subcutaneous expression	Adipose - Subcutaneous hmm	Adipose - Visceral (omentum) hmm	Distance to index	Liver hmm	Pancreatic acinar cell	Pancreatic beta cell	Pancreatic ductal cell	Pancreatic islet hmm	Pancreatic islet hmm	Pancreatic islet hmm
2.006	1.844	0.096	3	-0.061	253053	0	1.997	1.521	1.495	0	1	1.6
Downstream												
TP53INP1	TP53INP1_8_96092422	0.8	Features	Gene features summary								

91

Visualize the evidence



92

Now what...
Compare, Contrast & Learn



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

93

Combine to interpret & learn



DK Jang

Effector gene predictions

The ultimate goal of genetic association studies is to discover which genes and pathways have direct roles in risk of a disease or trait. Several methods now combine genetic association results with multiple kinds of evidence to generate lists of the genes that are most likely to mediate a genetic association. The currently available methods are described below, with links to their results.

T2D Effector Prediction Summary

View for:

Type 2 diabetes

Method:

In order to compare the results from different effector gene prediction methods, we have created the [T2D Effector Prediction Summary interface](#). This is a first attempt to summarize and integrate the results of these disparate methods.



94

Visualize & query the results



95

Visualize & query the results

Gene	Region	Top score	Curated Approach	EIP	ICP	View
	133161774					
HMGA1	6 : 34204650 - 34214008	5		5		Evidence
JADE2	5 : 133860003 - 133918920	5		5	1	Evidence
DNMT3A	2 : 25450724 - 25565459	5		5		Evidence
TP53INP1	8 : 95938200 - 95961639	5	2	4	5	Evidence
Curated Approach probability	Curated Approach locus	EPI probability	EPI locus	ICP probability	ICP locus	
POSSIBLE		0.860884726	8 : 95435147 - 96210768	0.8	TP53INP1_8_95685147	
JAZF1	7 : 27870192 - 28220362	5	3	5		Evidence
ZNF771	16 : 30418618 - 30442429	5	5			Evidence
WSCD2	12 : 108523248 - 108644314	5	5	3		Evidence
WFS1	4 : 6271576 - 6304992	5	5			Evidence
TM6SF2	19 : 19375173 - 19384200	5	5			Evidence
TBC1D4	13 : 75857639 - 76056250	5	5			Evidence
SLC5A1	22 : 32439019 - 32509016	5	5			Evidence
SLC30A8	8 : 117962512 - 118000000	5	5	4		Evidence

96

Now what...



- Compare, contrast, learn
- Enriching the predictions with the complete underlying data
- Updating the list as more data/evidence is generated



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

97

Integrating & making evidence accessible

Experiment summary for TSTSR043623
 Status: released ▲ 1

Summary		Attribution	
Assay:	HIC (Hi-C)	Lab:	Bing Ren, UCSD
Biosample summary:	<i>Homo sapiens</i> islet of Langerhans female adult (56 years) and female adult (59 years) and male adult (53 years)	Award:	5U01DK105541-03 (Bing Ren, UCSD)
Biosample Type:	tissue	Project:	AMP
Replication type:	anisogenic	References:	doi:10.1101/299388 PMID:31064983
Description:	HiC data from Bing Ren's Lab at UCSD with 3 non-diabetic individuals	Aliases:	bing-ren.HiC_isletOfLangerhans_3samples
Download Files:	Download	Date released:	2019-03-14

98

Current supporting data for the Curated T2D Effector Gene list

Genetic evidence- well captured

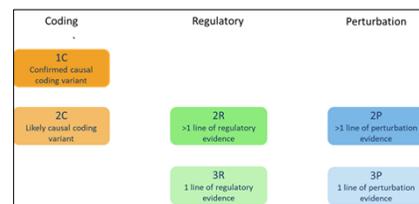
- GWAS, Credible sets & Gene-level association scores, Exome sequence results
- Monogenic associations from OMIM
- *Literature evidence from 11 papers for 11 genes*

Regulatory evidence- integration established, limited

- 11 genome-wide studies: 7 loaded in DGA, in progress for 4
- *Literature evidence from 13 papers for 15 genes*

Perturbation evidence- only summations

- *Literature evidence from 14 papers for 27 genes*
- Mutant phenotype annotations extracted from Zebrafish, Mouse, & Rat Genome Databases



ACCELERATING MEDICINES PARTNERSHIP (AMP)
 TYPE 2 DIABETES

99

Fully populating the current Curated Effector Gene list

Genetic evidence

- Monogenic associations from OMIM- [load the results to the portal](#)
- Literature evidence from 11 papers/genes- [load the results to the portal](#)



Maria Costanzo

Regulatory evidence

- Create direct links to the genome-wide annotations in DGA- [make accessible](#)
- Replace the literature citation & extract the gene-specific datasets from GEO, [load to DGA, curate a summary on the site](#)

Perturbational evidence

- [Extract mutant phenotype annotations](#) from Human Phenotype Ontology & annotations from Zebrafish, Mouse, Rat, *Drosophila*, *C. elegans* genome databases



ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

100

Updating & extending curated effector gene predictions

Genetic evidence- [expanded dataset, ancestries, traits](#)

- Add credible sets & gene-level association scores from large association studies for multiple phenotypes
- Update results with new GWAS M/As & for additional ancestries

Regulatory evidence- [iterative, better uptake in loading single gene studies](#)

- Identify large-scale studies for regulatory evidence types: eQTLs in relevant tissues; chromatin conformation in relevant tissues; allelic imbalance studies; others

Perturbation evidence-[integration of the resource into site](#)

- Extract mutant phenotype annotations from Human Phenotype Ontology & from Zebrafish, Mouse, Rat, *Drosophila*, *C. elegans* genome databases



ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

101

Onward...



- Curated Prediction from other disease communities
- Extending the approach(es) to all portal traits
- Unique ways to view genes prioritized by the portal data



ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

102

CARDIoGRAM Curated Effector Genes



CARDIoGRAMplusC4D (2021) effector gene list Type 2 diabetes

[View data](#) [View research method](#)

Near gene Evidence

*Click 'Evidence' button to view evidence data. *Hover evidence tables to see evidence group names.

Show all feature rows Hide feature headers Hide top level rows

14:100073487-100173487	LDLR	LPPR2,CCDC15,9,TMEM205	known	known	LDLR	6	LDLR	certain	closest gene; rare coding variants; PoPS; mouse phenotype	Evidence
		Number of enriched tissues		Strongest tissue						
		1		Adipose						
		95% credible set		Best SNP PPA		Rare variant association		Monogenic disorder		Missense coding variant
		10		0.34		LDLR		LDLR, LDLR-AS1		
		Top eQTL gene		Other eQTL genes		Top STARNET eQTL gene		Other STARNET eQTL genes		
		LDLR (BLD)		-		-		-		
		Mouse phenotype		UKBB PhenWAS Diseases		UKBB PhenWAS Continuous traits		PhenoScanner (non-UKBB)		
		LDLR		-		Apolipoprotein_B Cholesterol LDL_direct Triglycerides		-		
14:100073487-100173487	LDLR		known	known	LDLR	6	LDLR	certain	closest gene; rare coding	Evidence
6:161047871-161147871	LPL		known	known	LPL	6	LPL	certain	closest gene; coding variant; PoPS; mouse phenotype	Evidence
6:161061700-161161700	NOS3	AOC1,CDK5,GBX1,KCNH2,NOS3,SLC4A2,SMA,RCD3,ASIC3,AB	known	known	NOS3	6	NOS3	certain	closest gene; PoPS; eQTL; mouse phenotype	Evidence



Krishna Aragam



103

Onward...

- GIANT Gene Predictions
- Implementing EI portal-wide
- Modeling the Mahajan Heuristic on additional traits
- Connecting with other efforts (ICDA, etc.)
- Expanding with AMP-CMD!



ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

104

Additional views within the portal



Emphasize relationships rather than genomic coordinates



105

Advancing prioritizations tools

- Decision support for your prior
- Building principled priors
- Future directions

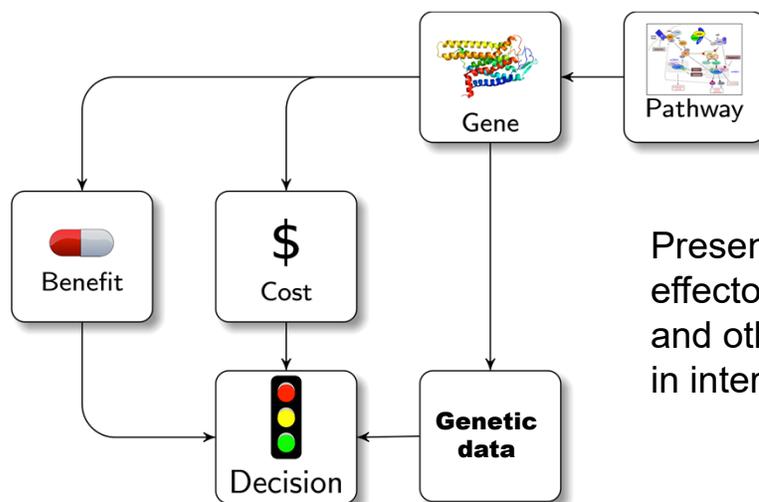


ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

106

Onward...



Present integrated views of effector gene predictions, and other data in the portal, in interpretable ways



ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

107

Gap 1: human genetic data is (or was) not *accessible*

- The T2D Knowledge portal allows users to interrogate a gene across hundreds of traits through a simple web interface

hugeamp.org

ACCELERATING MEDICINES PARTNERSHIP (AMP)

CMDKP

Home Data Tools Information Contact Login

COMMON METABOLIC DISEASES KNOWLEDGE PORTAL

Providing data and tools to promote understanding and treatment of common metabolic diseases

Gene, region or variant Phenotypes Disease-specific portals

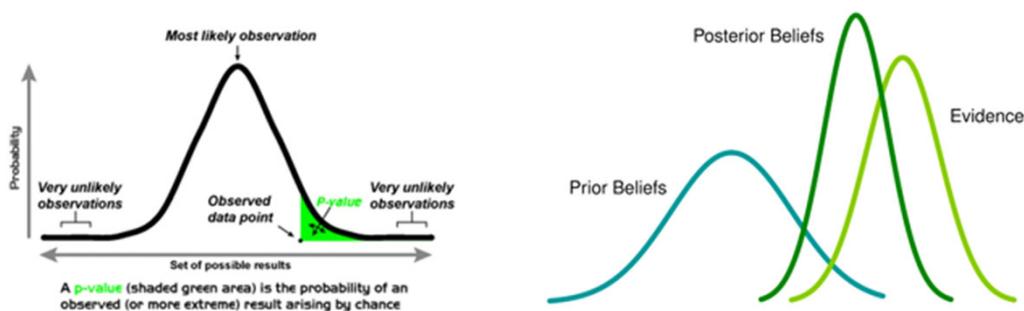
Search

examples: PCSK9, rs1260326, chr9:21,940,000-22,190,000

108

Gap 2: human genetic data is *hard to interpret*

- P-values do not tell you how likely an association is to be true
- Bayesian “posteriors” do measure this quantity, by updating a prior belief in light of evidence (e.g. the observed association)



109

Gap 3: what prior should we use?

- Most researchers don't have intuition for **quantitative** priors
- But we use **qualitative** priors all the time
- Are you more likely to believe in a T2D-susceptibility gene if...

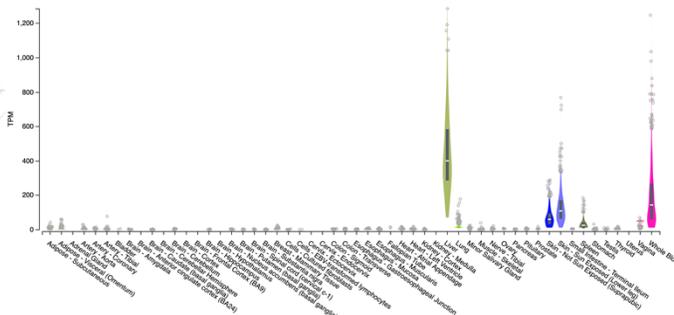
...it is specifically expressed in liver?

...a mouse knockout has dysglycemia?

Mouse Phenotypes

Availability	Mouse Genotype
Find Mice	Madg ^{tm1Bpra} /Madg ^{tm1Bpra} Tg(Ins2-cre/ERT)1Dam/0 (conditional)

decreased insulin secretion
increased fasting circulating glu.
hyperglycemia
decreased circulating insulin (fasting)
impaired glucose tolerance



110

Membership in a gene set affects a gene's prior

218 genotypes with 218 annotations displayed of selected term and subterms

Searched Term: [decreased insulin secretion](#)

Allelic Composition (Genetic Background)	Annotated Term	Reference
Abcc8^{tm1.1Fmas}/Abcc8^{tm1.1Fmas} (B6.129S2(Cg)-Abcc8 ^{tm1.1Fmas})	decreased insulin secretion	1:208932
Abcc8^{tm1.1Mgn}/Abcc8^{tm1.1Mgn} (involves: 129X1/SvJ * C57BL/6)	decreased insulin secretion	1:79352
Abcc8^{tm1.2brv}/Abcc8^{tm1.2brv} (involves: 129X1/SvJ)	decreased insulin secretion	1:179577
Abcc8^{tm1.2brv}/Abcc8^{tm1.2brv} Sstr5^{tm1Fcb}/Sstr5^{tm1Fcb} (involves: 129X1/SvJ)	decreased insulin secretion	1:179577
Abhdg^{tm1a(EUCOMM)Hmqv}/Abhdg^{tm1a(EUCOMM)Hmqv} (C57BL/6N-Abhdg ^{tm1a(EUCOMM)Hmqv})	decreased insulin secretion	1:234876
Ace^{tm3Unc}/Ace* (involves: 129P2/OlaHsd * C57BL/6J)	decreased insulin secretion	1:128859
Adipog^{tm1.1th}/Adipog^{tm1.1th} (involves: 129S7/SvEvBrd * C57BL/6J)	decreased insulin secretion	1:199264
Afm1g^{tm1b(EUCOMM)Wtsi}/Afm1g^{tm1b(EUCOMM)Wtsi} (C57BL/6N-Afm1g ^{tm1b(EUCOMM)Wtsi} /H)	decreased insulin secretion	1:226537
Akap5^{tm1.1jaco}/Akap5^{tm1.1jaco} (involves: 129S * C57BL/6)	decreased insulin secretion	1:190017
Akap5^{tm1Mdas}/Akap5^{tm1Mdas}	decreased insulin secretion	1:190017

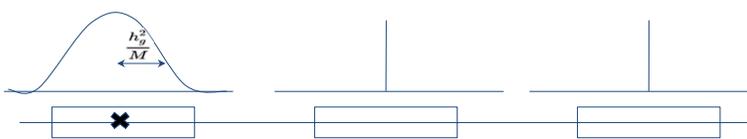
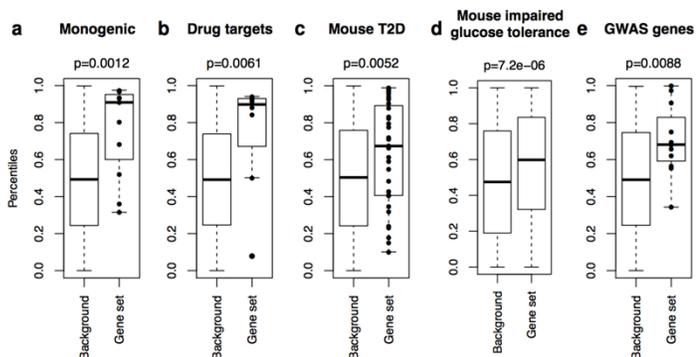
111

The degree to which should be measurable as a gene set enrichment

- Either:
 - Stronger than expected **rare variant** associations

and/or

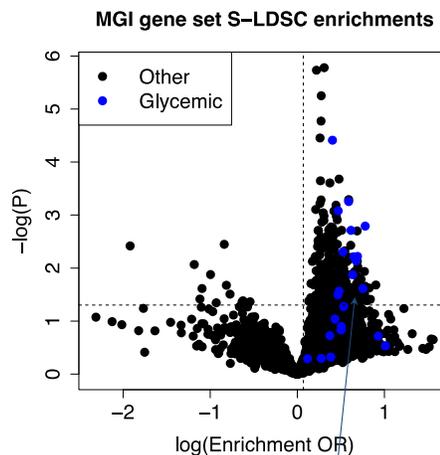
- Excess heritability explained by **common variants** nearby the genes



112

Application to “mouse knockout” gene sets

- Calculate heritability enrichments (via Stratified LD-score regression) for each of ~3000 mouse knockout gene sets
- Enrichment odds-ratio roughly corresponds to a “relative prior”



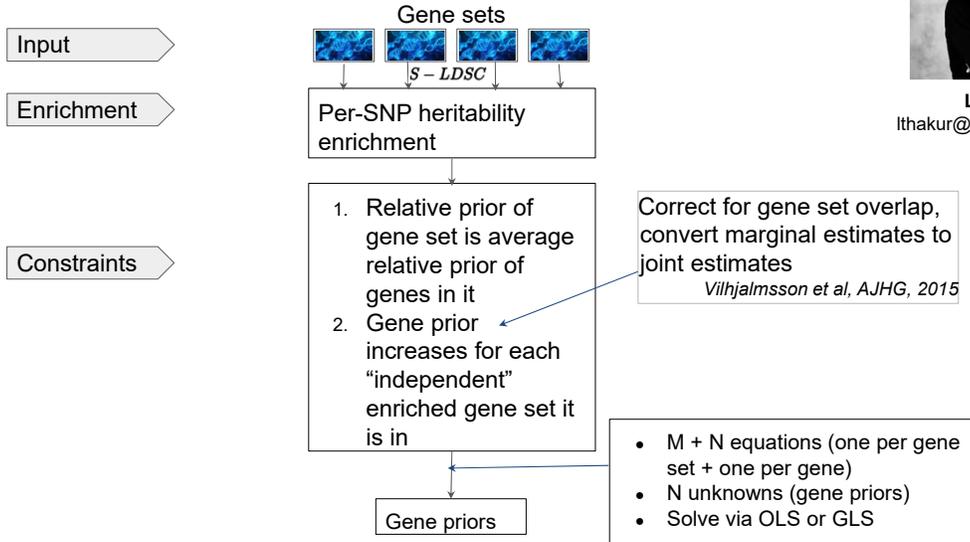
24/24 glycemic phenotypes are positive (binomial $p=1.1 \times 10^{-4}$)

113

How to combine relative priors from multiple gene sets

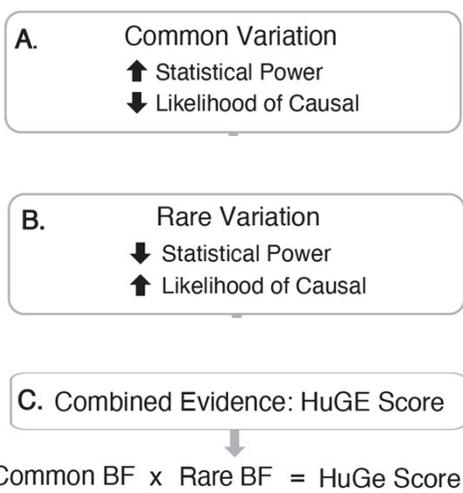


Lokendra Thakur
lthakur@broadinstitute.org



114

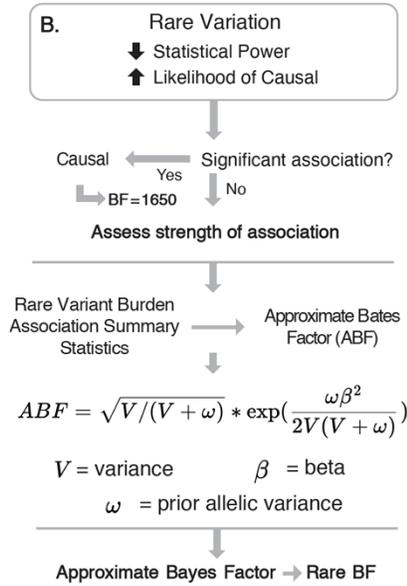
Gap 4: how does genetic data **update** our prior?



Peter Dornbos
pdornbos@broadinstitute.org

115

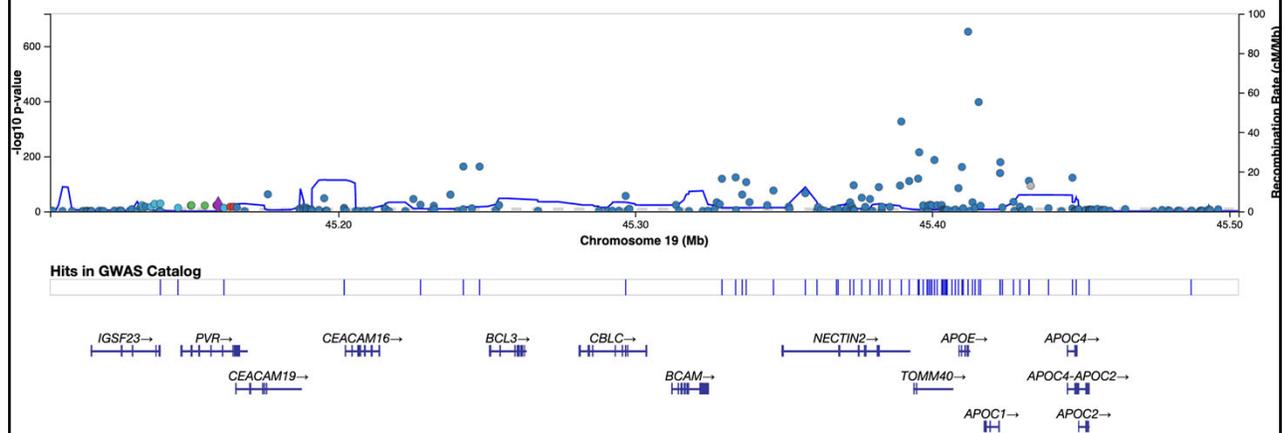
Rare variation



116

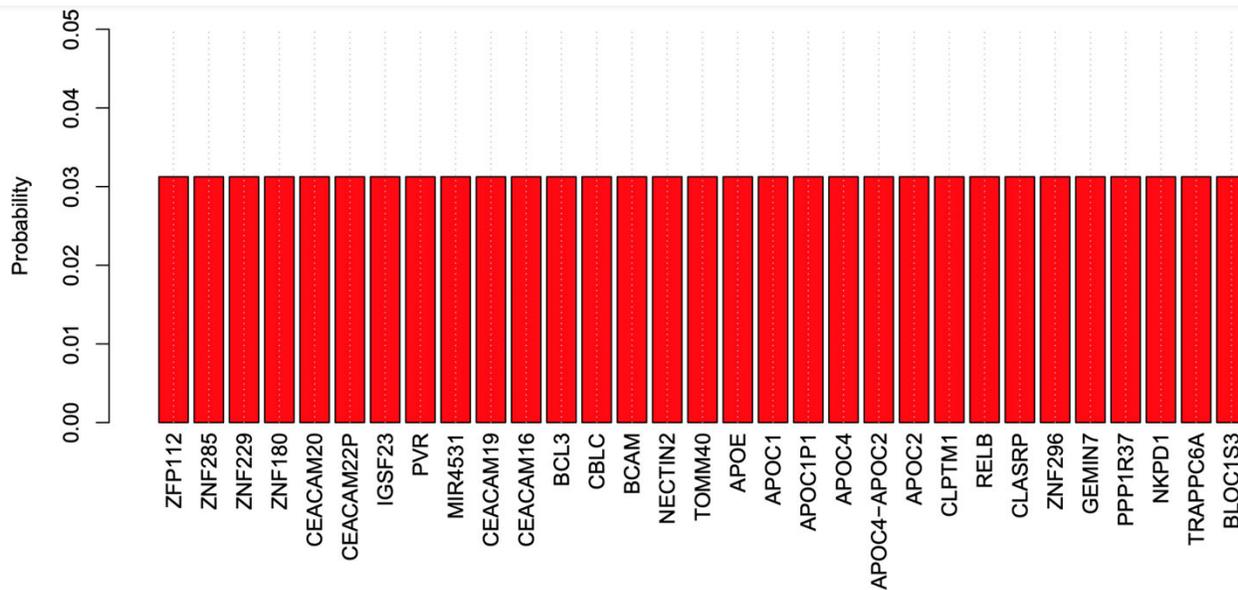
Common variation

- First question: is the gene in a GWAS region?



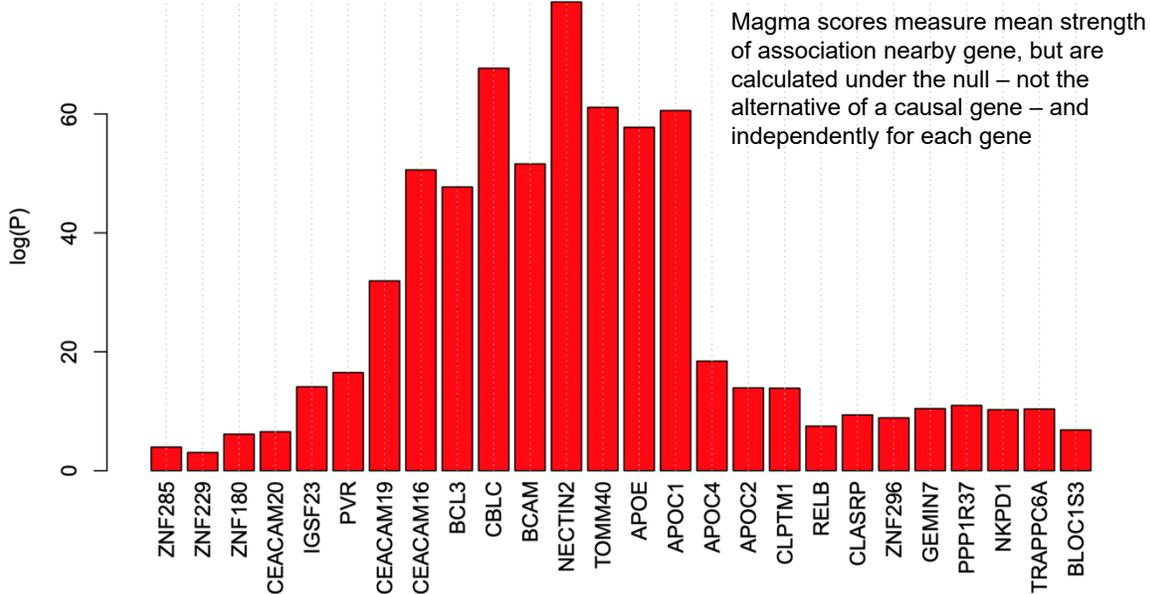
117

Naïve model: uniform likelihood



118

Less naïve model: MAGMA



119

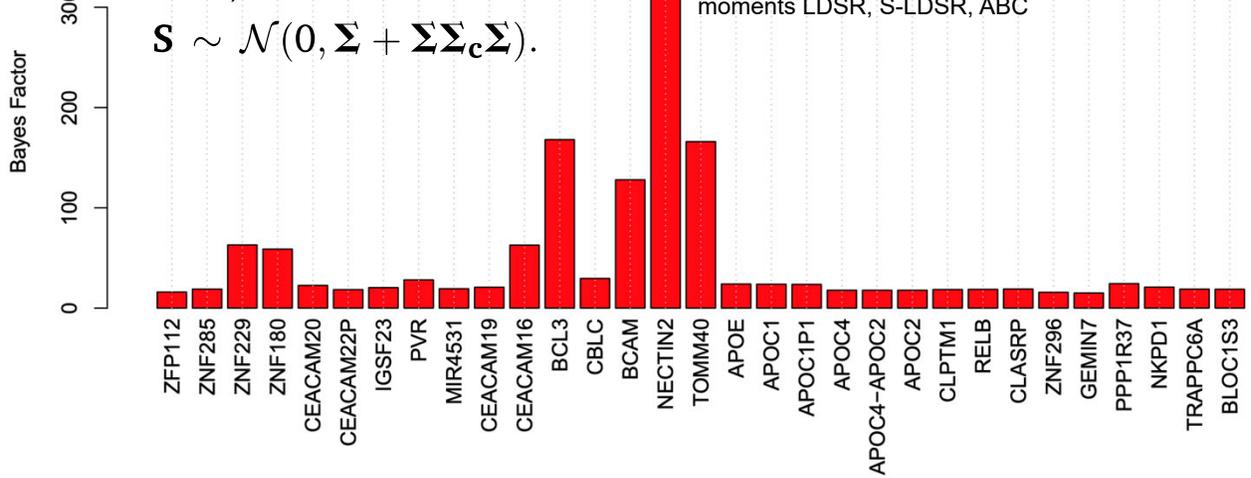
Under development: a Bayesian model

Causal SNPs inflate summary statistics for all SNPs in LD (idea underlying LD-score regression and fine mapping approaches such as CAVIAR):

$$S \sim \mathcal{N}(0, \Sigma + \Sigma \Sigma_c \Sigma).$$

Hypothesis: causal **genes** should increase number of causal SNPs nearby

Parameters for this model can in principle be estimated from available data such as LDSR, 4th moments LDSR, S-LDSR, ABC



120

Common variation



- Second question: is it the effector gene?

Gene	Region	Top score	Curated Approach	EIP	ICP	View
PCDH17	13 : 58205944 - 58303445	5			5	Evidence
DPEP1	16 : 89679716 - 89704864	5			5	Evidence
SLC12A8	3 : 124801480 - 124931708	5		5		Evidence
ATXN7	3 : 63884075 - 63989129	5		5	1	Evidence
BCL2	18 : 60790579 - 60987361	5		5		Evidence
FBRSL1	12 : 133066137 - 133161774	5		5		Evidence
HMGA1	6 : 34204650 - 34214008	5		5		Evidence
JADE2	5 : 133860003 - 133918920	5		5	1	Evidence
DNMT3A	2 : 25450724 - 25565459	5		5		Evidence
TP53INP1	8 : 95938200 - 95961639	5	2	4	5	Evidence
JAZF1	7 : 27870192 - 28220362	5	3	5		Evidence
ZNF771	16 : 30418618 - 30442429	5	5			Evidence
WSCD2	12 : 108523248 - 108644314	5	5	3		Evidence
WFS1	4 : 6271576 - 6304992	5	5			Evidence
TM6SF2	19 : 19375173 - 19384200	5	5			Evidence
TBC1D4	13 : 75857639 - 76056250	5	5			Evidence
SLC5A1	22 : 32439019 - 32509016	5	5			Evidence

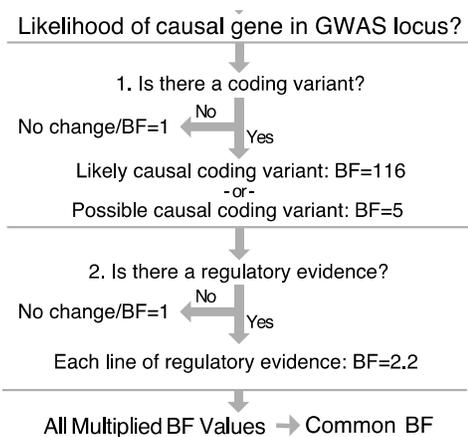
ACCELERATING MEDICINES PARTNERSHIP (AMP)

TYPE 2 DIABETES

121

Converting effector gene predictions to probabilities

- Some methods directly output probabilities
- Others are qualitative – but, with some transparent assumptions, we can convert them to probabilities
- For the curated list, assuming...
 - “Causal” genes are 95% confident
 - “Strong” genes are 80% confident
 - 2C/2R evidence lines are equivalent
 - Lines of evidence are independent
 Then...



122

Putting it all together: The HuGE calculator



Preeti Singh
psingh@broadinstitute.org

Gene: INSR Search gene Phenotype: Type 2 Diabetes

Combined Evidence ⓘ
INSR has Moderate evidence of a disease-susceptibility.

Causal Strong Moderate Possible Potential Weak No Evidence

Common Variation ⓘ
Common variation evidence for INSR in Type 2 Diabetes...

Genetic Evidence: 2C

Causal Strong Moderate Possible Potential Weak No Evidence

INSR is Genome-wide significant.

Rare Variation ⓘ
Rare variation evidence for INSR in Type 2 Diabetes

Causal Strong Moderate Possible Potential Weak No Evidence

INSR is not Exome-wide significant.

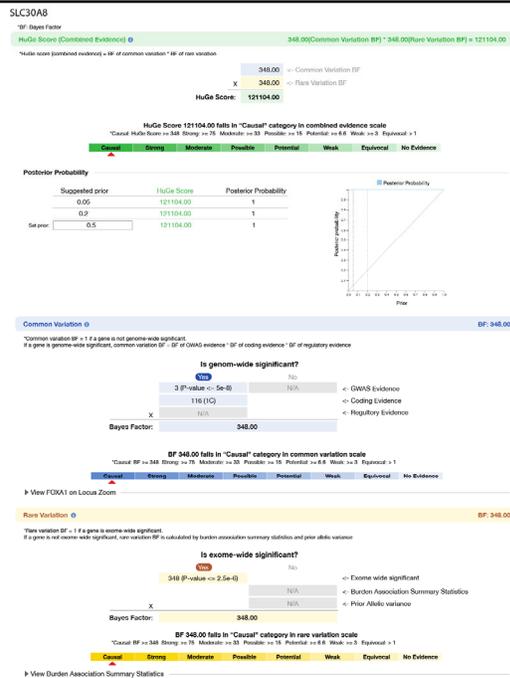
INSR is not Exome-wide significant since its p-value is greater than the threshold p-value of 2.5e-6. So Bayesian approach is used to calculate PPAs over a range of priors. By default the prior variance is 0.3696 as shown in the line plot below. Please

Posterior Probability

123

Long term vision

- Present a “gene focused” entry into the effector gene list
- And (over time as our model improves) a gene focused entry into all portal data



124

The Multi-Site Team

Broad/DCC
 Kenneth Bruskiwicz
 Lizz Caulkins
 Maria Costanzo
 Marc Duby
 Clint Gilbert
 Quy Hoang
 DK Jang
 Ryan Koesterer
 Jeffrey Massung
 Oliver Ruebenacker
 Preeti Singh

Leadership
 Jason Flannick
 Noël Burt
 Mike Boehnke
 Kyle Gaulton
 Thomas Keane
 Jose Florez



125

PANEL DISCUSSION

