Effector Gene Prediction Symposium

AMP T2D Consortium

May 21, 2021

FNIH
Foundation for the National Institutes of Health

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** Nöel Burtt and Jason Flannick
1:20 p.m.	7.	**PANEL DISCUSSION**
1:50 p.m.		*Meeting Adjourn*

FNIH

Accelerating visualization and target prediction tools through AMP T2D

Phil Smith and Melissa Thomas
AMP T2D Co-Chairs

FNIH
Foundation for the National Institutes of Health

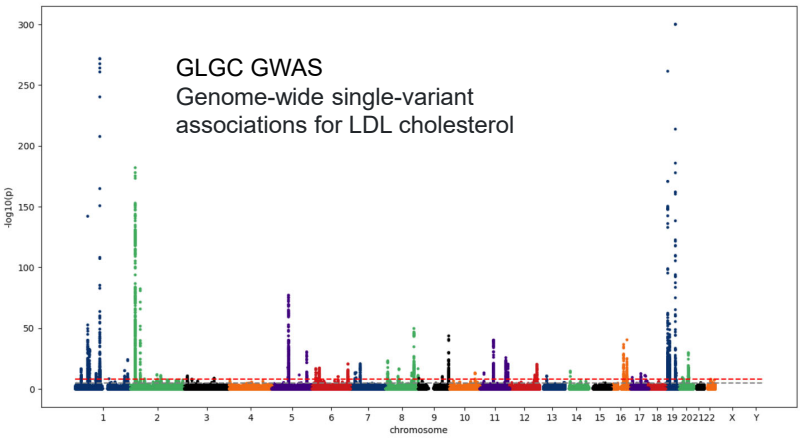
Considerations for target prioritization tools

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

May 21st, 2021

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:

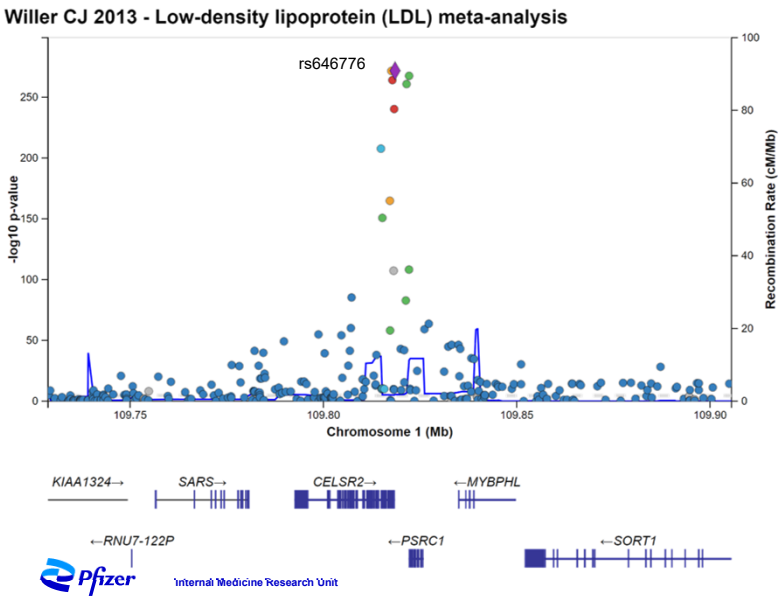
Why does this gene influence this trait?

and

How does the genetic variation impact the gene?

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

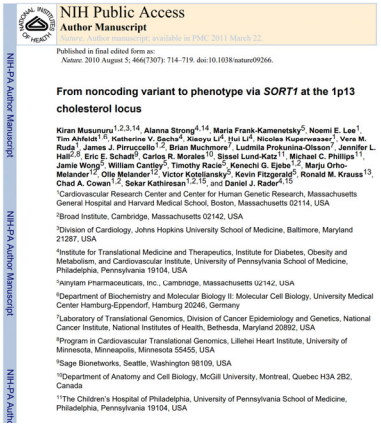


Lead variant	rs646776
Causal or functional variant	?
Impact of variant on effector transcript	?
Causal gene/effector transcript	?
Function or activity of effector transcript	?
Measured phenotype	**LDL-C levels**

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
Nature. 2010 August 5; 466(7307): 714–719. doi:10.1038/nature09266.

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

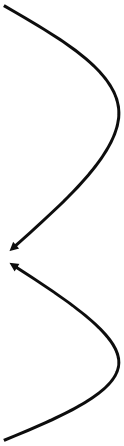
Musunuru et al, Nature. 2010 466(7307):714–719

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

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"

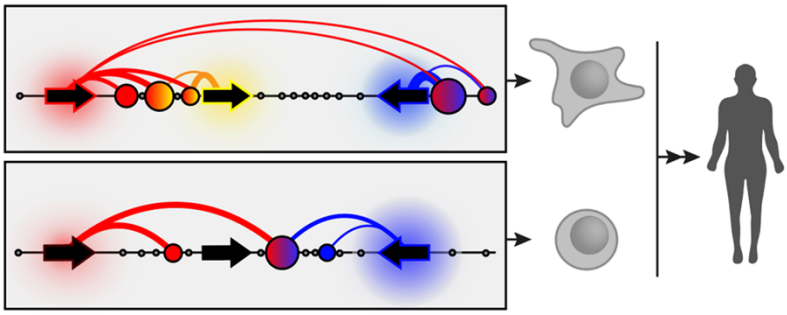
What gene is contributing to this trait?
Which gene should we go after?

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

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

Brent Richards

FNIH
Foundation for the National Institutes of Health



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

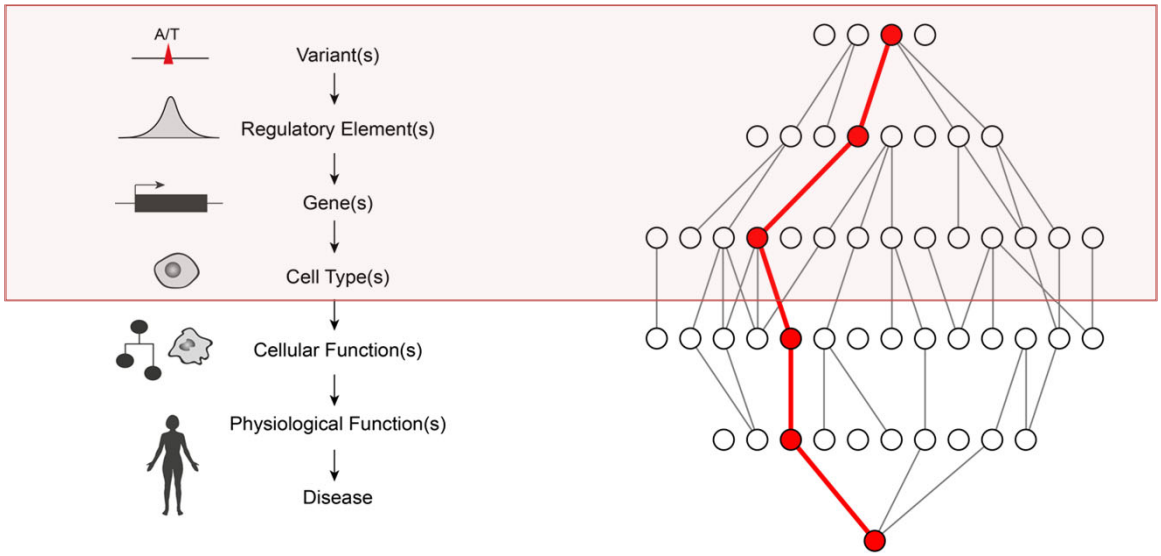
Jesse Engreitz, Ph.D.
BASE Initiative, Moore Children's Heart Center
Department of Genetics, Stanford University
Broad Institute of MIT and Harvard

12

No disclosures to report

13

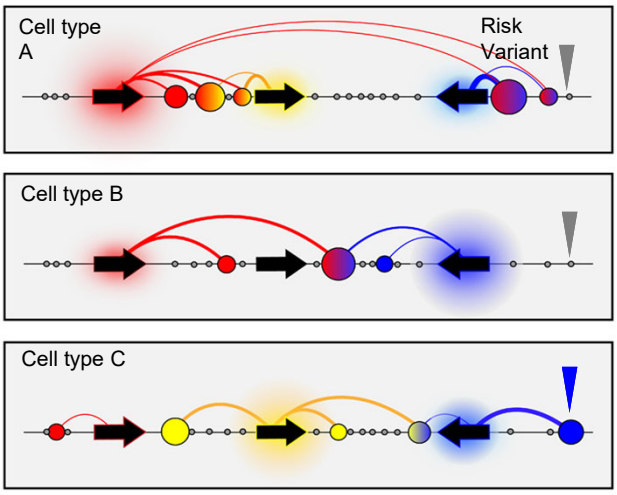
Connecting all T2D variants to enhancers and cell types



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

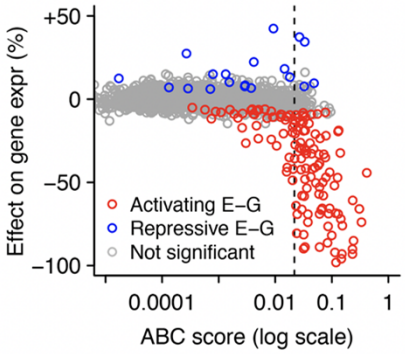
15

A simple formula predicts E-G connections

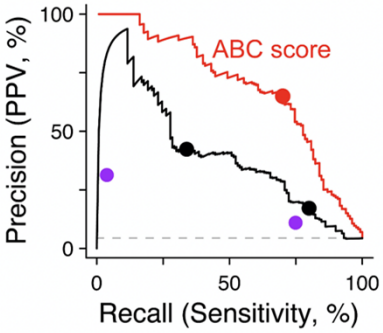
$$\text{Predicted effect of enhancer on gene} = \frac{\text{Activity} \times \text{Contact}}{\Sigma\,(\text{Activity} \times \text{Contact})}$$

Activity = DHS × H3K27ac
Contact = Hi-C contact frequency

Predict quantitative effects:

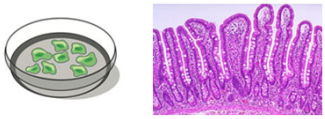


Classify regulatory E-G pairs:

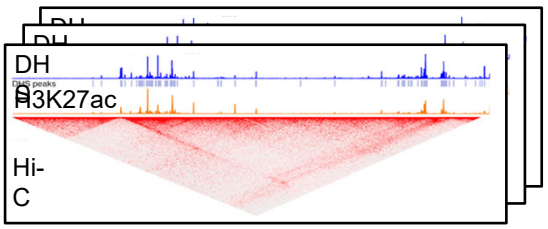


16

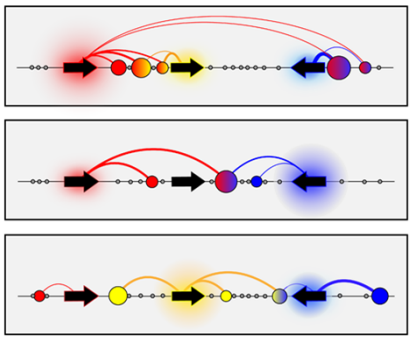
ABC predicts enhancer-gene connections across cell types



Epigenome maps



ABC

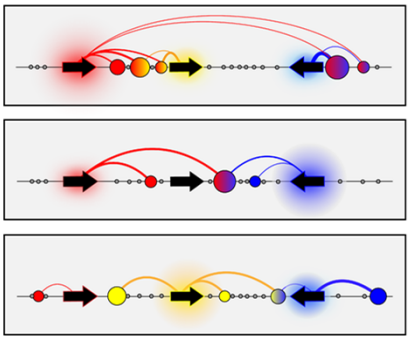


Enhancer-promoter maps

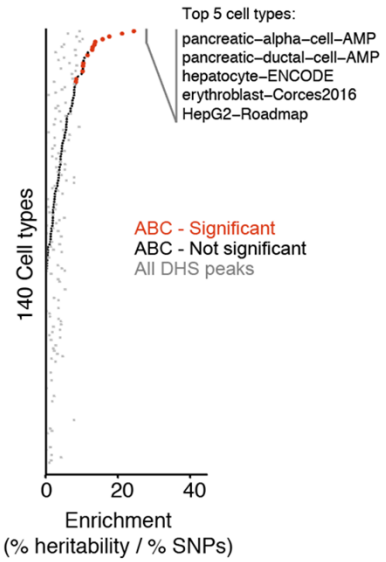
We can now create draft regulatory maps of the noncoding genome from epigenomic data across many cell types

17

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



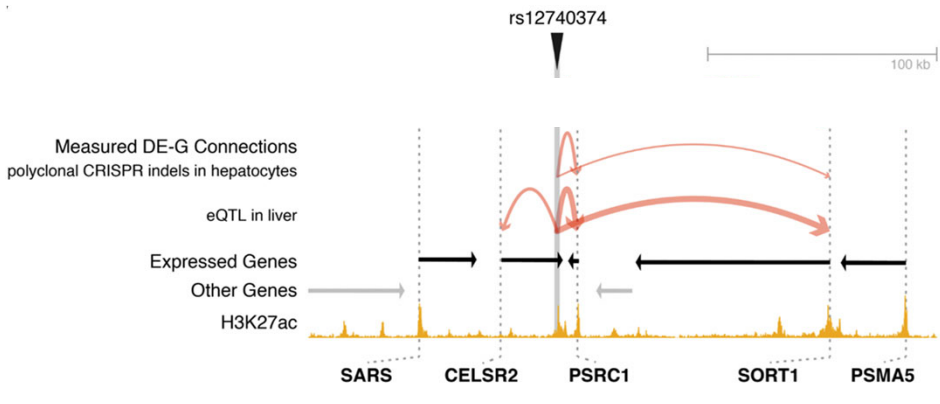
AMP-T2D:
Pancreatic cells, adipocytes, hepatocytes, muscle



Melina Claussnitzer, Alisa Manning, Tim Majarian

18

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



Musunuru *et al.*
Nature 2010

SORT1

↓

Hepatocytes

↓

PCSK9
trafficking

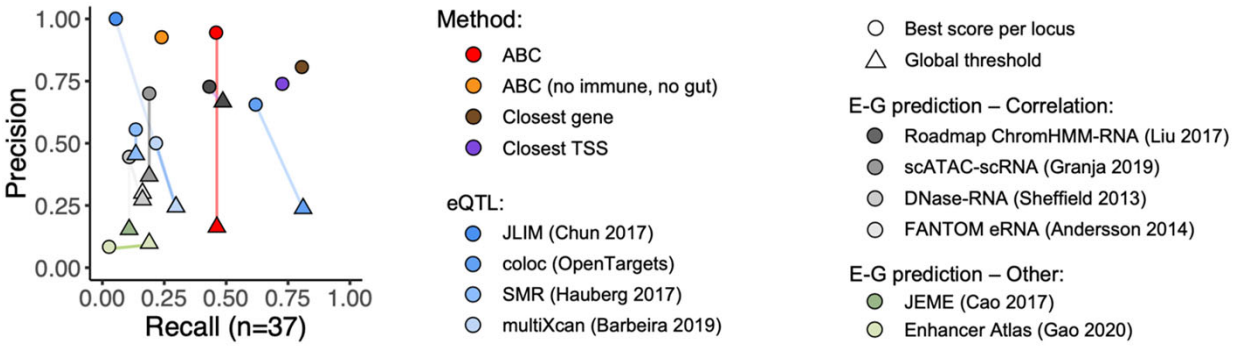
↓

LDL +
Atherosclerosis

Fulco *et al.* Nat Genet 2019

19

Set 1: IBD genes

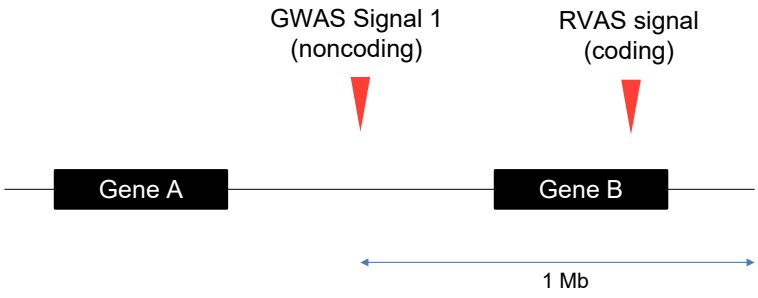


IBD Known Genes: Curated by domain expert based on prior evidence from coding variation, knowledge of gene function, mouse models

Nasser *et al.* Nature 2021

Can we collect other sets of "silver-standard" genes?

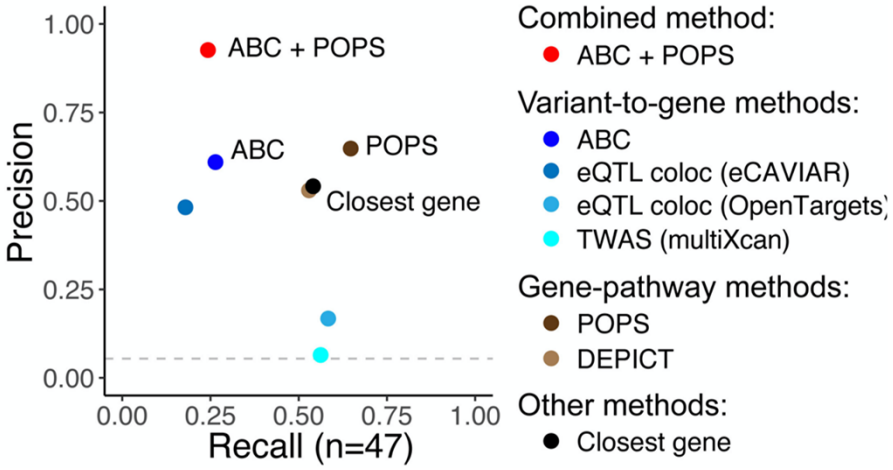
Idea: Use genes containing rare coding variants as positive controls



with Elle Weeks, Hilary Finucane

Set 2: LDL cholesterol

LDL cholesterol: Genes carrying rare coding variants from RVAS

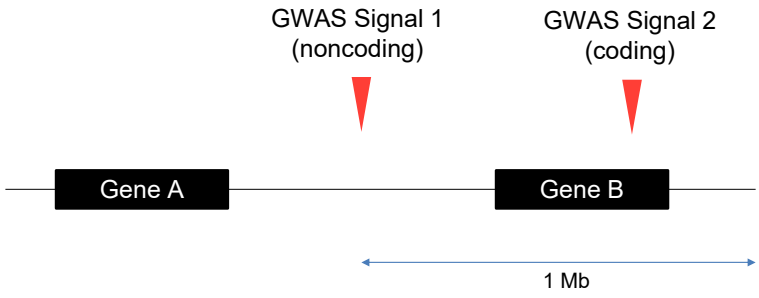


with Hilary Finucane, Elle Weeks

22

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

Idea: Use genes containing coding variants as positive controls



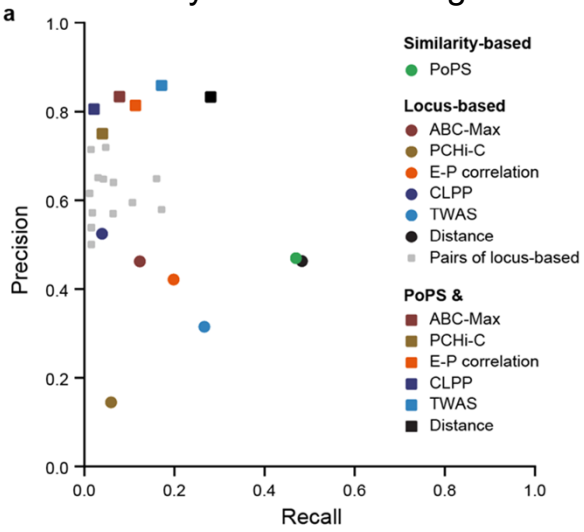
There are >600 such examples across 73 UK Biobank traits

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

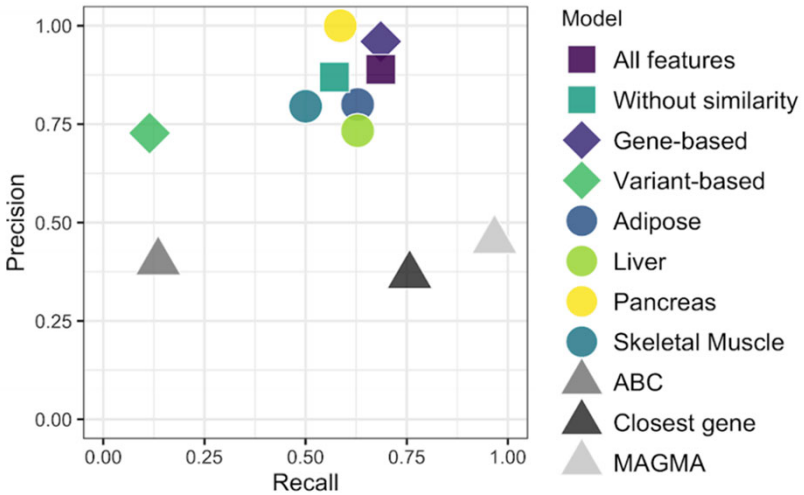


Weeks *et al.*, medRxiv 2020

24

Set 4. Type 2 diabetes

"Causal" effector genes from T2DKP



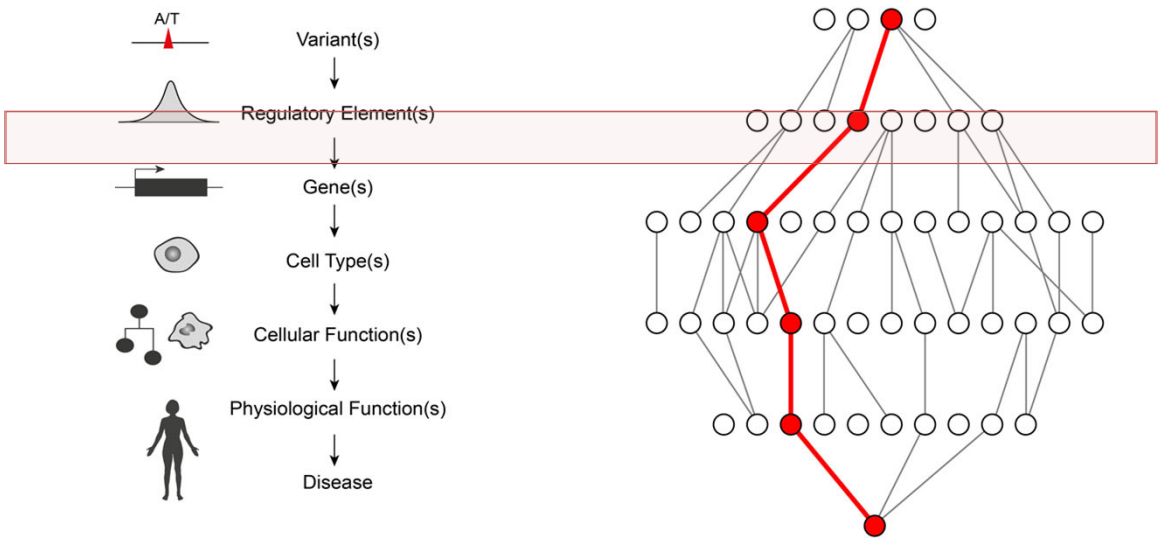
Tim Majarian, Alisa Manning

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?

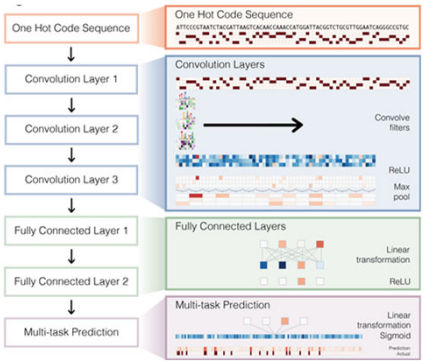
Connecting all T2D variants to enhancers and cell types



Connecting risk variants to enhancers

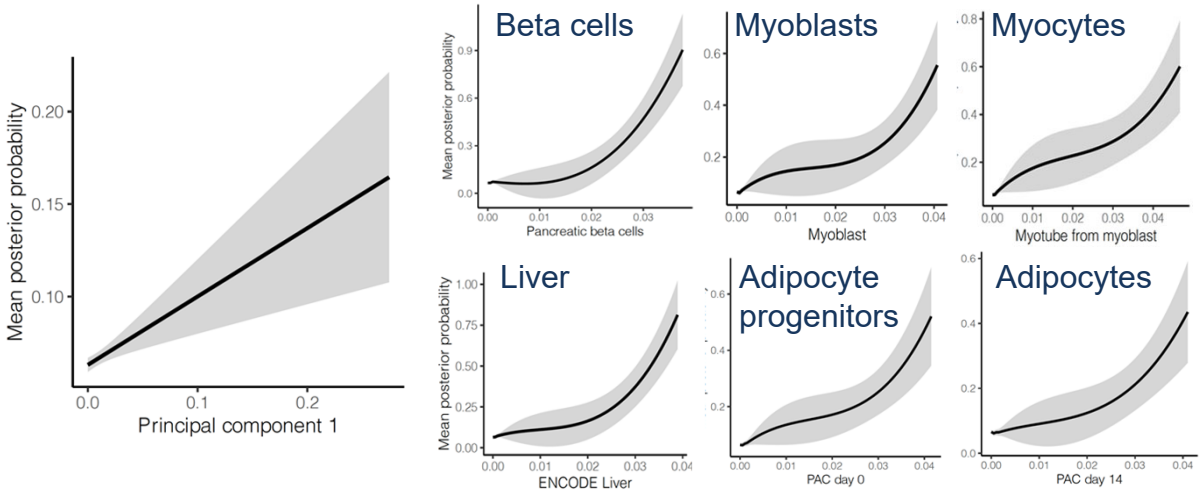
Predict V-E combo for all T2D variants from sequence data using CNN scores

- Trained a deep learning model to predict chromatin data from DNA sequence
- 117 islet, adipose, muscle, and liver ChIP-seq, DNase-seq, and ATAC-seq datasets (tissue and cell types)
- Model achieves AUC > 0.8 for the validation dataset



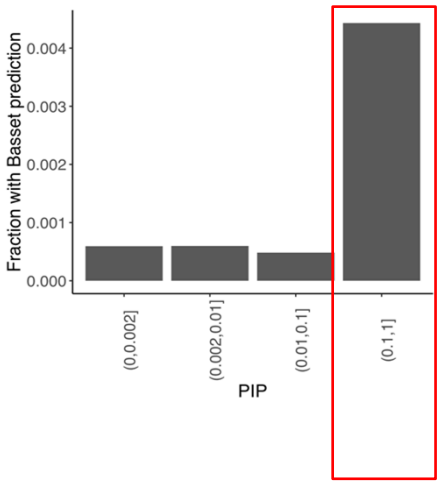
Connecting risk variants to enhancers

Validate sequence-based scoring scheme using posterior probabilities

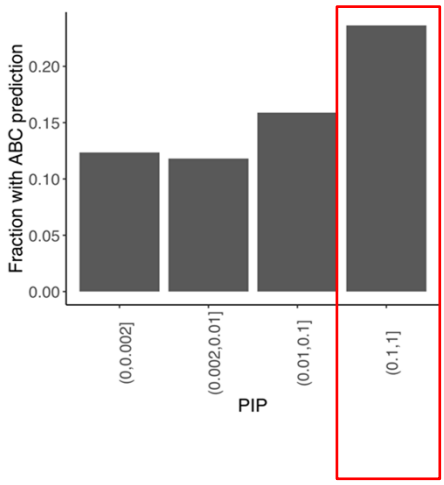


Connecting risk variants to enhancers to genes

CNN variant predictions (Basset score >= 0.03)

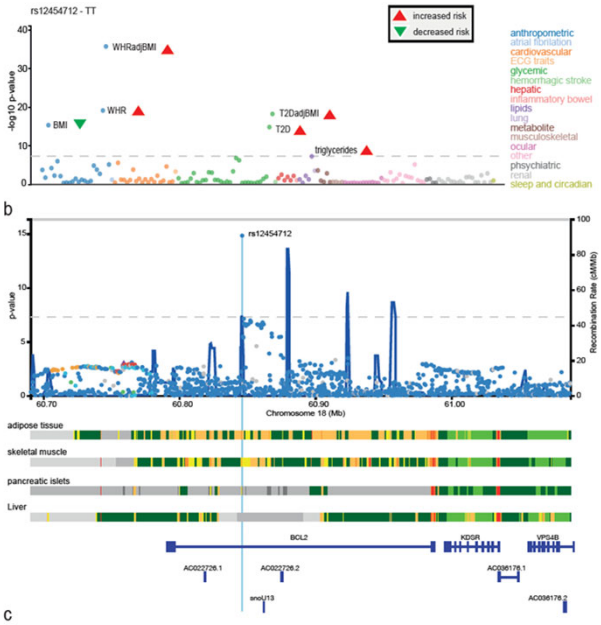


ABC V-G predictions

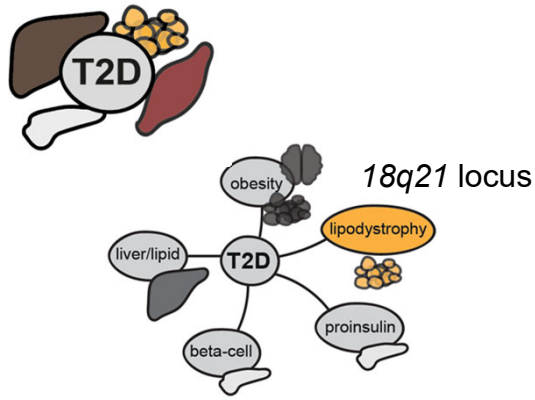


30

Connecting V2F for the *18q21* metabolic risk locus



Metabolically unhealthy lean phenotype



18q21 locus

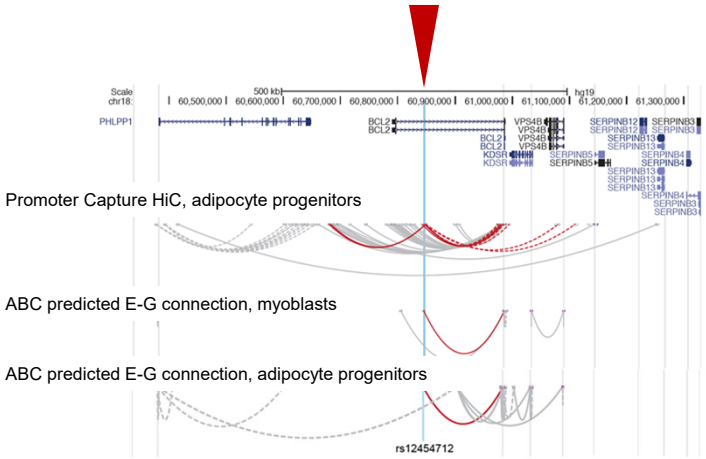
Udler et al, 2018

31

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

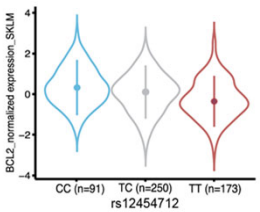
ABC effector gene predictions

rs12454712

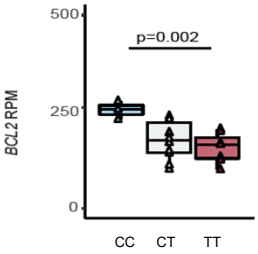


eQTLs

Skleletal muscle (STARNET database)



Adipocyte progenitors

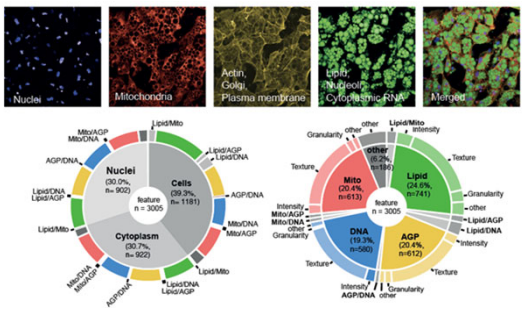


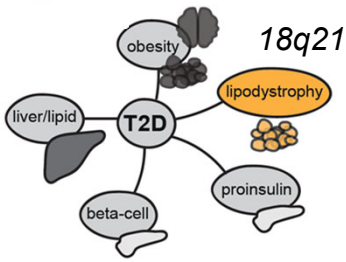
32

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





18q21 locus

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

33

Connecting V2F for the *18q21* metabolic risk locus

LipocyteProfiler

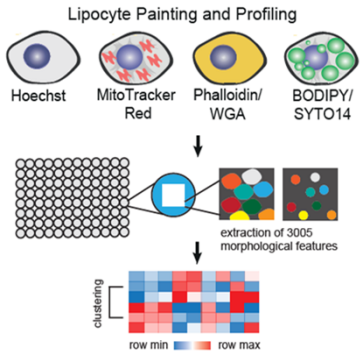


Image-based profiling to resolve depot-specific effects

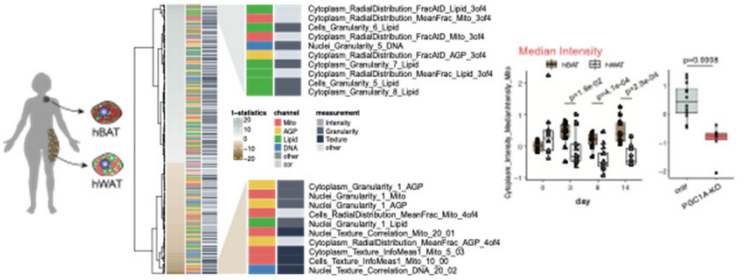
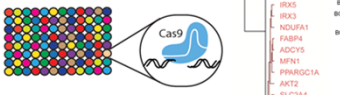
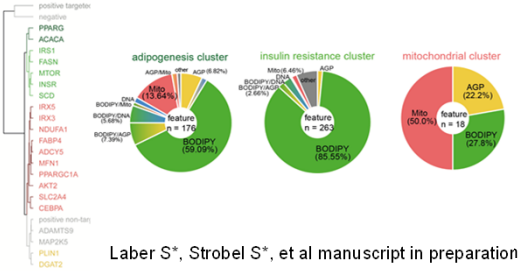


Image-based profiling to classify marker genes into pathways



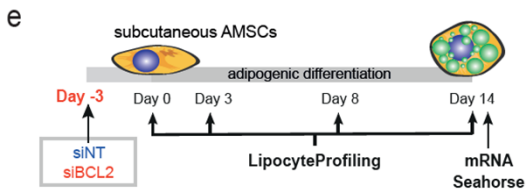


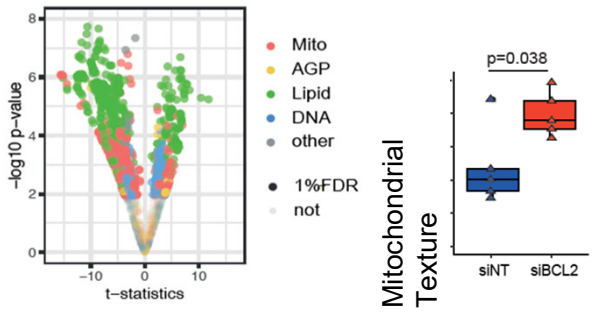
Laber S*, Strobel S*, et al manuscript in preparation

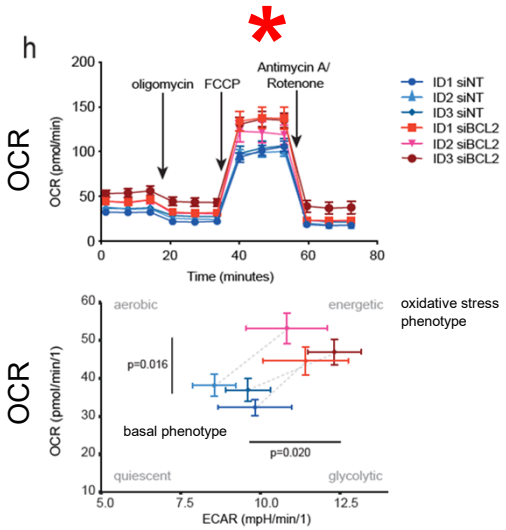
34

Connecting G2F for the *18q21* metabolic risk locus

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





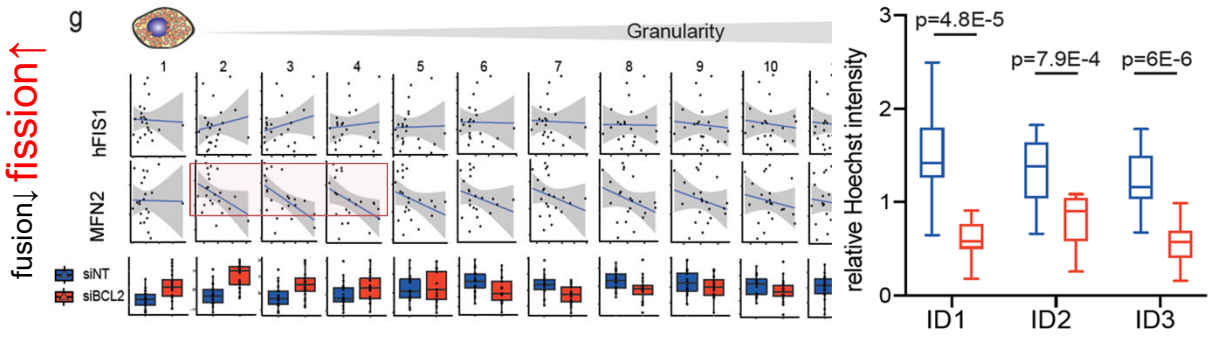


Laber S*, Strobel S*, et al manuscript in preparation

35

Connecting G2F for the *18q21* metabolic risk locus

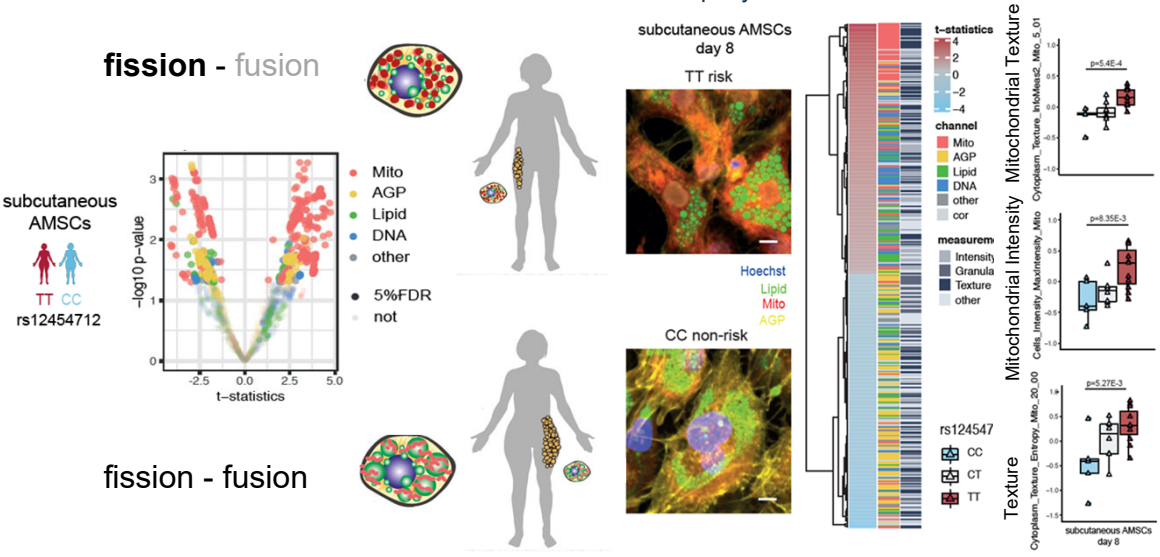
Directed *BCL2* perturbation in subcutaneous adipocyte progenitors results in increased mitochondrial fission and apoptosis



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

Connecting V2F for the *18q21* metabolic risk locus

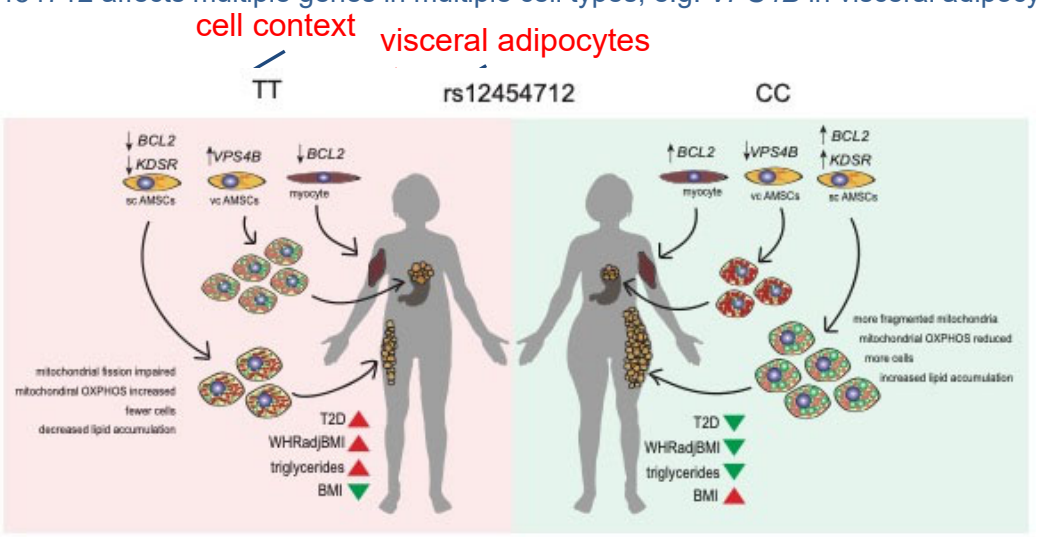
rs12454712 controls mitochondrial fission in subcutaneous adipocytes via *BCL2*



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

Connecting V2F for the *18q21* metabolic risk locus

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

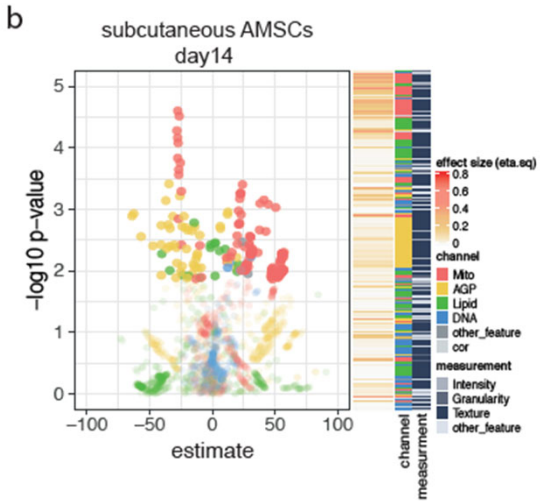
Mark Daly
Hailiang Huang

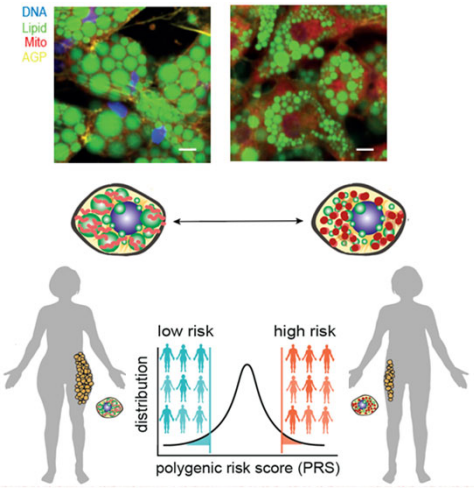
Chuck Epstein

BROAD INSTITUTE
HARVARD UNIVERSITY
NIH National Human Genome Research Institute
INTERNATIONAL COMMON DISEASE ALLIANCE | Maps to Mechanisms to Medicine

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

FNIH
Foundation for the National Institutes of Health

43

Outline

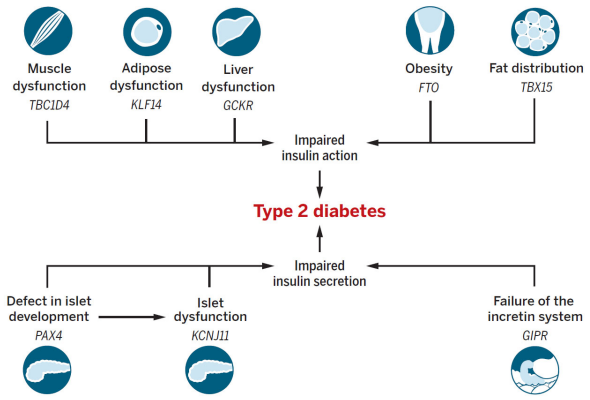
- What are the things that we know that a model should include?
- How did we come up with the heuristic model, and what did we find?
- Are there any updates given subsequent data?
  - additional GWAS (DIAMANTE, MVP)
  - additional genomic data (islets, fat, muscle, liver…)
- How does the heuristic model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



FNIH

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.

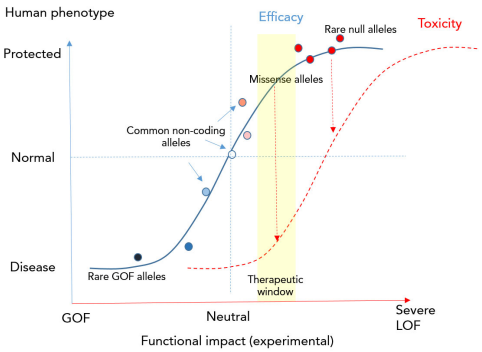
This really helps to locate the regulatory (or other) mechanisms through which signals operate, and (for regulatory variants) can home in on the cell-type, state and process involved

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

FNIH

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:

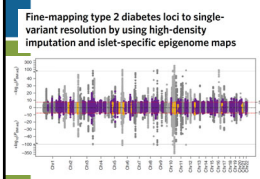
- Not every credible set coding variant is causal (Mahajan NG 2018a)
- The "equivalence" assumption may be violated: the phenotypic spectrum of coding & regulatory variants for the same gene may differ (due to tissue expression)
- The "proximity" assumption (ie that multiple signals at a locus operate 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

FNIH

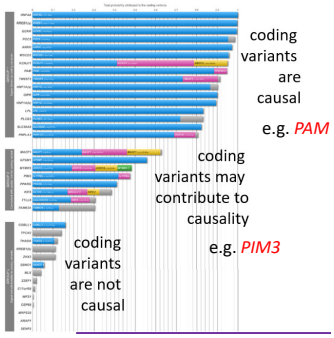
47

Causal coding variants

GWAS | Exome array | Exome sequencing | Rare variant linkage



e.g. *CDKN1B*
PATJ



Single variant
e.g. *PAX4*

Gene-based
e.g. *SLC30A8*

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

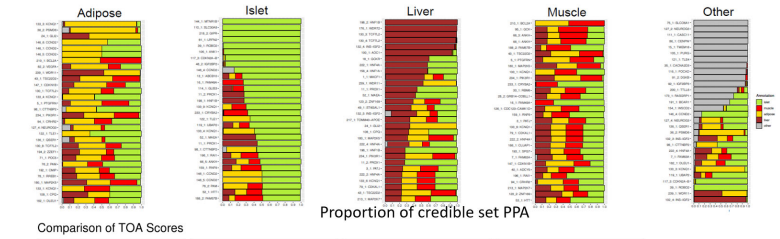
What do we know? (4)

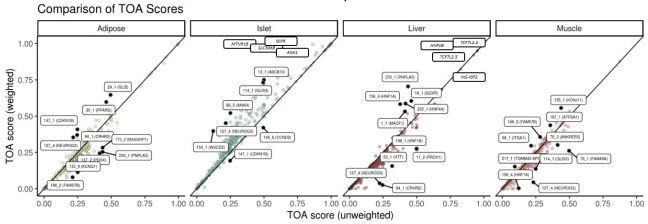
Indeed, as density of signals gets ever higher, the concept of a "locus" is increasingly redundant

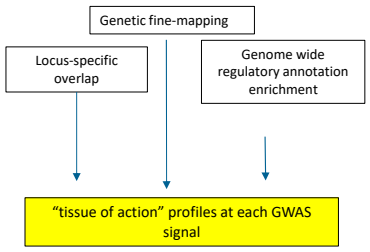
- Some loci involve 20+ signals (*INS/IGF2/KCNQ1*) – are all of these working through the same mechanism?
- Not all signals at the same locus influence the same cell type (*TCF7L2*)
- We have clear examples where there are two overlapping signals with distinct mechanisms (NKX6.3 via islets; ANK1 via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

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







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

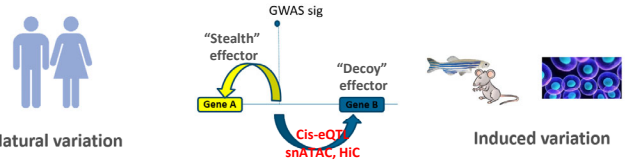
- Sensitivity 67%
- Specificity 8% (☊ to 19% if co-localizing)
- Lots of "bystander" genes indistinguishable from the true gene

- 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)

FNIH

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.



Natural variation

Coding v associations
Monogenic genes
Human KO alleles

Induced variation

Animal models
Cellular studies

CRISPR….

FNIH

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….

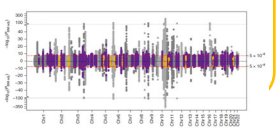
FNIH

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

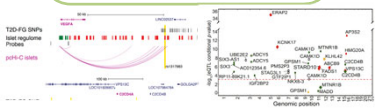
Mahajan et al Nature Genetics 2018a
Mahajan et al Nature Genetics 2018b
Flannick et al Nature 2019
OMIM



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 Genetics 2019
Viñuela et al Nat Commun 2021



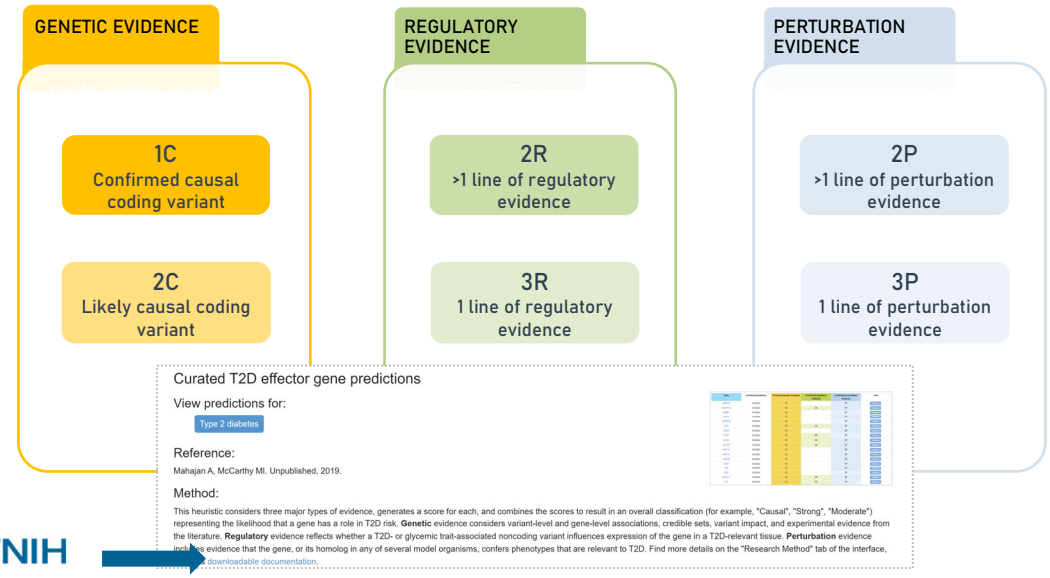
PERTURBATION EVIDENCE

- 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

FNIH

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	Previously associated loci
GLP1R	

GWAS coding evidence	Exome array evidence	Burden test evidence	Monogenetic associations	Other genetic evidence
	Strong	Medium		PMID:27252175 | PMID:29941447

Islet cis-eQTLs	Other relvant cis-eQTLs	Islet chromatin conformation	Allelic imbalance	Glucose regulation	Other regulatory evidence
				1	

RNA interference evidence	Zebrafish mutant phenotype	Mouse mutant phenotype	Drosophila mutant phenotype	Rat mutant phenotype	Other perturbation evidence
		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		decreased circulating glucose level | increased heart rate | increased systemic arterial blood pressure | increased insulin secretion	

Gene	Combined prediction	Combined genetic evidence	Combined regulatory evidence	Combined perturbation evidence	View
HNF1A	CAUSAL	1C		2P	Evidence

Predicted T2D effector gene	Previously associated loci
HNF1A	

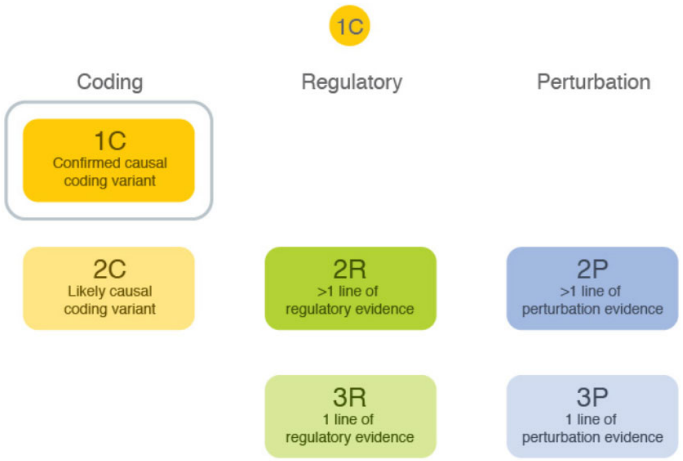
GWAS coding evidence	Exome array evidence	Burden test evidence	Monogenetic associations	Other genetic evidence
Strong	Strong	Strong	MODY	PMID:24915262

Islet cis-eQTLs	Other relvant cis-eQTLs	Islet chromatin conformation	Allelic imbalance	Glucose regulation	Other regulatory evidence
				0	PMID:24915262

RNA interference evidence	Zebrafish mutant phenotype	Mouse mutant phenotype	Drosophila mutant phenotype	Rat mutant phenotype	Other perturbation evidence
increased insulin secretion	pancreatic B cell decreased amount | pancreatic B cell decreased area + glucose homeostasis disrupted	abnormal liver morphology | abnormal 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			

56

Causal



N=41

Primary GWAS coding eg *SLC30A8*

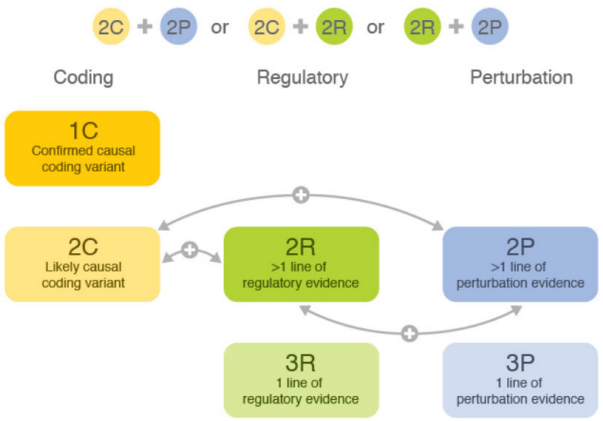
Secondary GWAS coding eg *HNF1A*

Exome sequencing eg *PAX4*

Monogenic nearby eg *WFS1, PPARG*

57

Strong



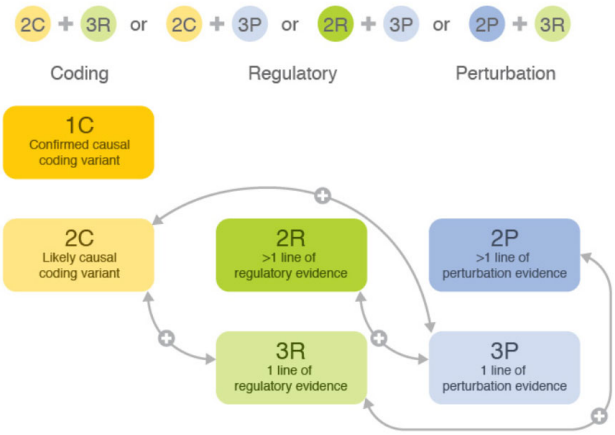
N=17

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

IRS1, CAMK1D, KLF14, TCF7L2

58

Moderate



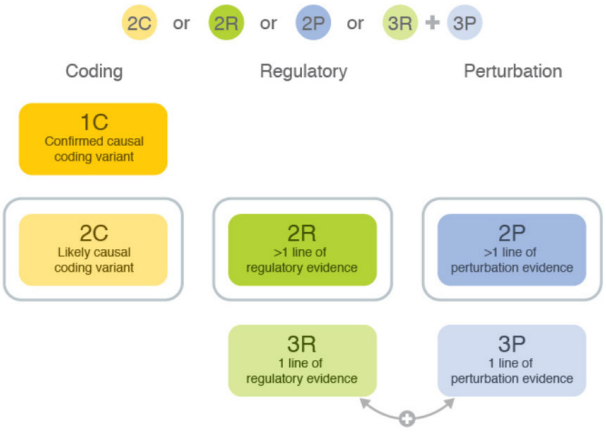
N=21

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

Eg *ADCY5, IGF2BP2, JAZF1*

59

Possible



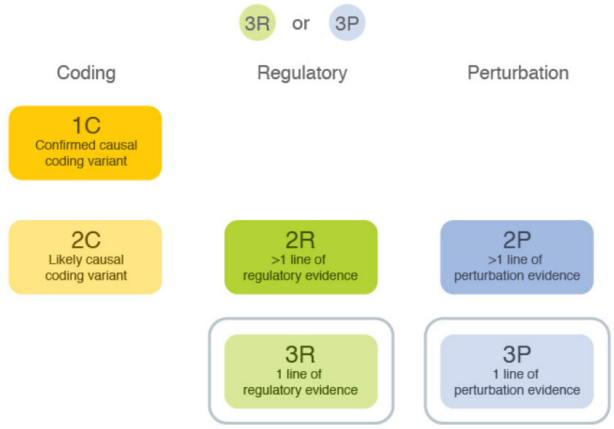
N=13

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

e.g. *ST6GAL1*

60

Weak



N=6

One piece of evidence

e.g. *ITGB6*

61

T2D-related

N=34

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

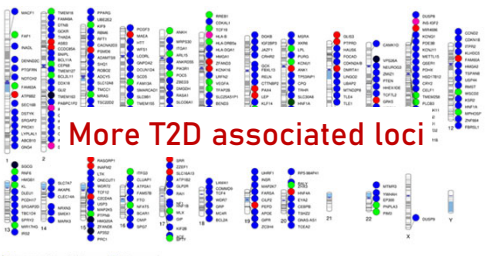
- **Likely causal for monogenic diabetes or related trait eg *CEL, AKT2***
- **Evidence connecting to variation in continuous glycemic traits *eg GRB10, SIX2/3***

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

62

What has changed since then?

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



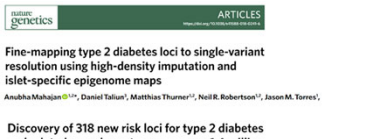
More T2D associated loci

European ● South Asian ● Latino
East Asian ● African American ● Multi-ethnic

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*

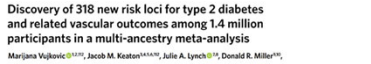


Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps

Anubha Mahajan, Daniel Taliun, Matthias Thurner, Neil R. Robertson, Jason M. Torres,

Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis

Marijana Vujkovic, Jacob M. Keaton, Julie A. Lynch, Donald R. Miller,



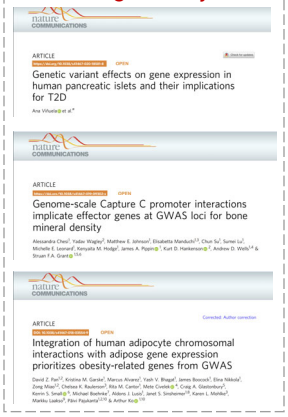
Identification of type 2 diabetes loci in 433,540 East Asian individuals

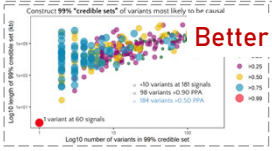
Cassandra N. Spracklen, Momoko Horikoshi, Young Jin Kim, Kuang Lin,

Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation

Anubha Mahajan, Cassandra N Spracklen, Weihua Zhang, Maggie CY Ng, Lauren E Petty, Hideoshi Kitajima, Grace Z Yu, Sina Rüeger, Leo Speidel, Young Jin Kim, Momoko Horikoshi, Josep M Mercader, Daniel Taliun,

More regulatory data





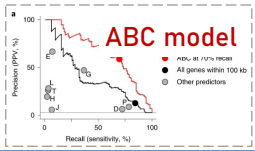
Better fine-mapping resolution

More exomes

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

PETER DORNBOS, LAURA RAFFIELD, XIANYONG YIN and JASON FLANNICK

Tissue of action

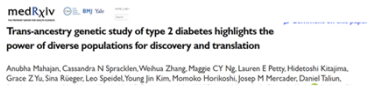


ABC model

Expanded candidate effector transcript list

Coding: 19 signals including *MYO5C* & *ACVR1C*

Protein QTLs: PAM, PGM1, PRSS3, APOE (7 signals at 7 regions)



Trans-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation

Anubha Mahajan, Cassandra N Spracklen, Weihua Zhang, Maggie CY Ng, Lauren E Petty, Hideoshi Kitajima, Grace Z Yu, Sina Rüeger, Leo Speidel, Young Jin Kim, Momoko Horikoshi, Josep M Mercader, Daniel Taliun,

RNA QTLs

126 cis-eqtl at 66 signals

Islets (42)	SAT (28)	VAT (18)	SM (23)	Liver 8	Brain 3
CEP68	*CEP68*	*CEP68*	*CEP68*	*CEP68*	
STARD10	*IRS1*	*IRS1*	*ANK1*	*JAZF1*	*CLAUP1*
DGKB	*KLF14*	*INHBB*	*ZZEF1*	*SLC22A3*	
NKX6.3	*ANK1*	*PCGF3*	*PCGF3*		
ST6GAL1	*PLEKH1*	*JAZF1*	*JAZF1*		
	PRC1		*SKOR1*		

Chromatin Interactions

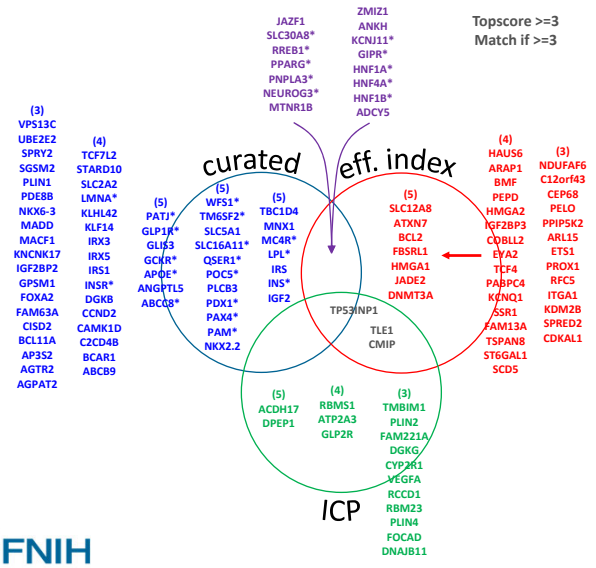
214 contacts

Islets	SAT	Liver
TCF7L2	*CCND2*	*CMIP*
RFX3	*IRX5*	*CCND2*
ST6GAL1	*GIPR*	*PPARG*
CTRB2	*PPARG*	*SLC22A3*
KCNK17	*TCF7L2*	*FOXA2*
	ANK1	

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

FNIH

Comparisons from the portal



	curated	EI	ICP
1: 40Mb	MACF1	PABPC4	
2: 65Mb		CEP68 SPRED2	
3: 185Mb	IGF2BP2	ST6GAL1	DGKG DNAJB11
5: 102Mb	PAM	PPIP5K2	
6: 7Mb	RREB1	RREB1 SSR	
6: 39Mb	GLP1R KCNK17		
7:23Mb		IGF2BP3	FAM221A
8: 95Mb		TP53INP1 NDUFAF6	TP53INP1
9:19Mb		HAUS6	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	

FNIH

65

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)

FNIH

66

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

FNIH

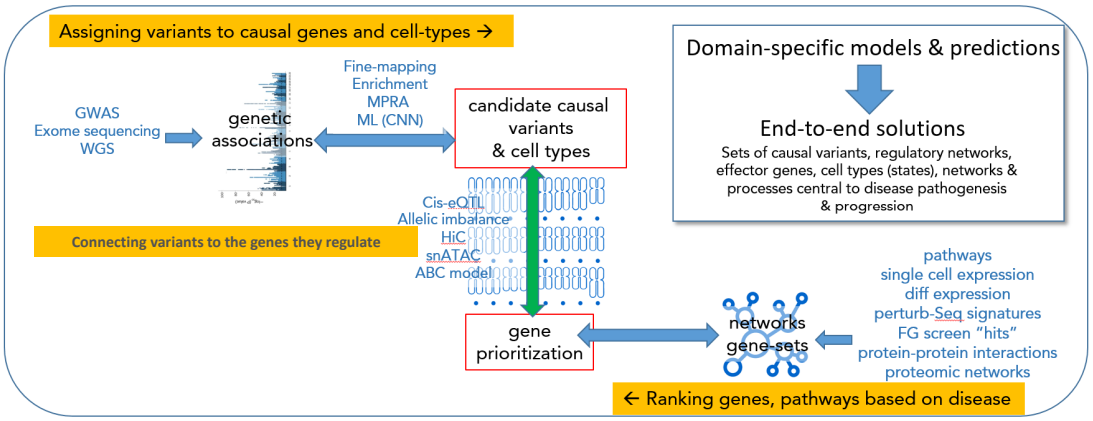
Can we do better?

- **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)
- **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

Allelic series: dose-response

$$S_{1..i} = S_1 + \frac{S_2}{2^2} + \frac{S_3}{3^2} + \frac{S_4}{4^2} \ldots + \frac{S_i}{i^2}$$

Next steps

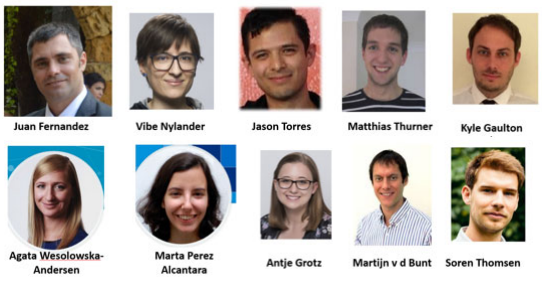


Weeks et al – POPS
Jagadeesh et al – sc and GWAS
Fernandez-Tajes et al – GWAS & PPI data

FNIH

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
- Maike Sander
- Bing Ren
- Patrick MacDonald
- Jesse Engreitz
- Jason Flannick
- Manolis Dermitzakis
- Eric Fauman
- Robert Plenge
- Sally John

wellcome trust | NIH National Institutes of Health | FNIH Foundation for the National Institutes of Health | imi

ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

70

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

Noël Burtt & Jason Flannick
May 21, 2021

ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

72

Today

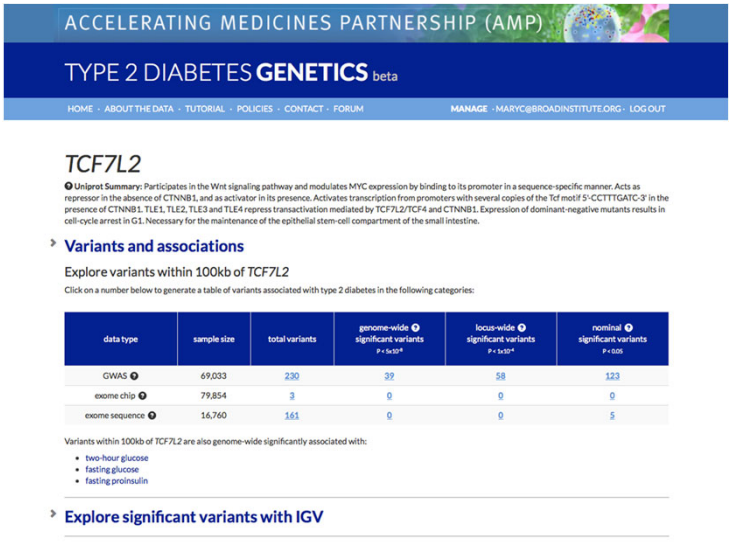
- Representation of effector prediction approaches
- Enhancing & expanding effector prediction approaches
- Decision support & advancing tools for visualization for gene & target prioritization
- Future directions

ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

73

Evolution of the portal

Dataset-level genetic association results

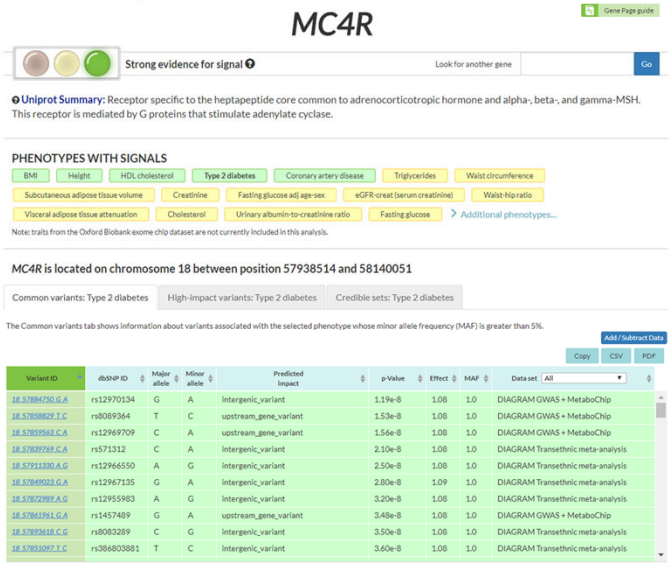


2015

74

Evolution of the portal

Distilled gene-level results

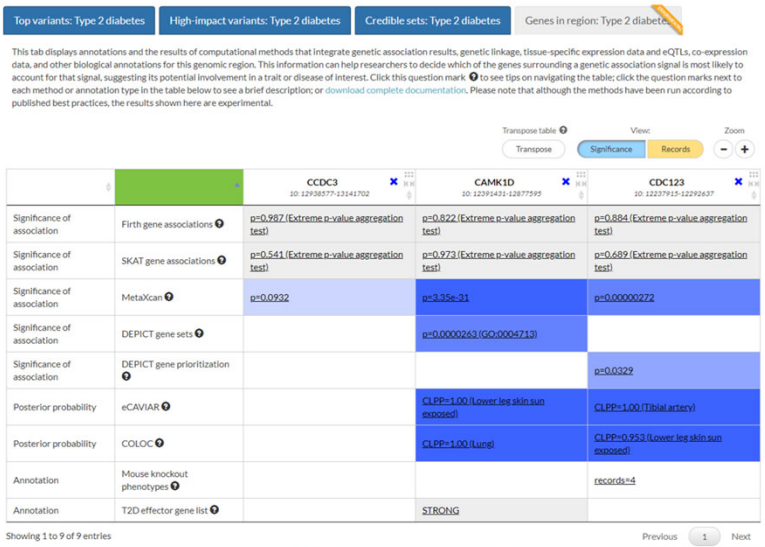


2017

75

Evolution of the portal

Results of methods to support gene prioritization



2019

76

Converting to Knowledge

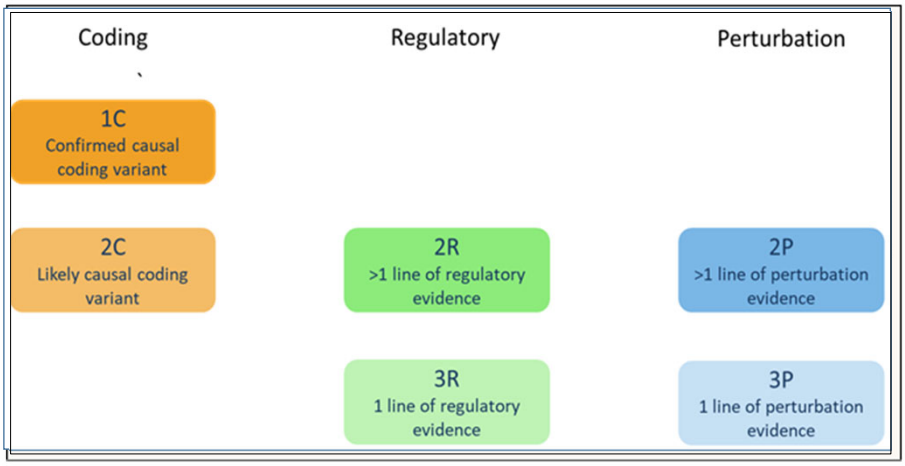
Building Effector Gene Knowledge ⟷ Tools for Prioritization

77

Curated Effector Transcript Heuristic


Anubha Mahajan


Mark McCarthy



1 Coding Evidence

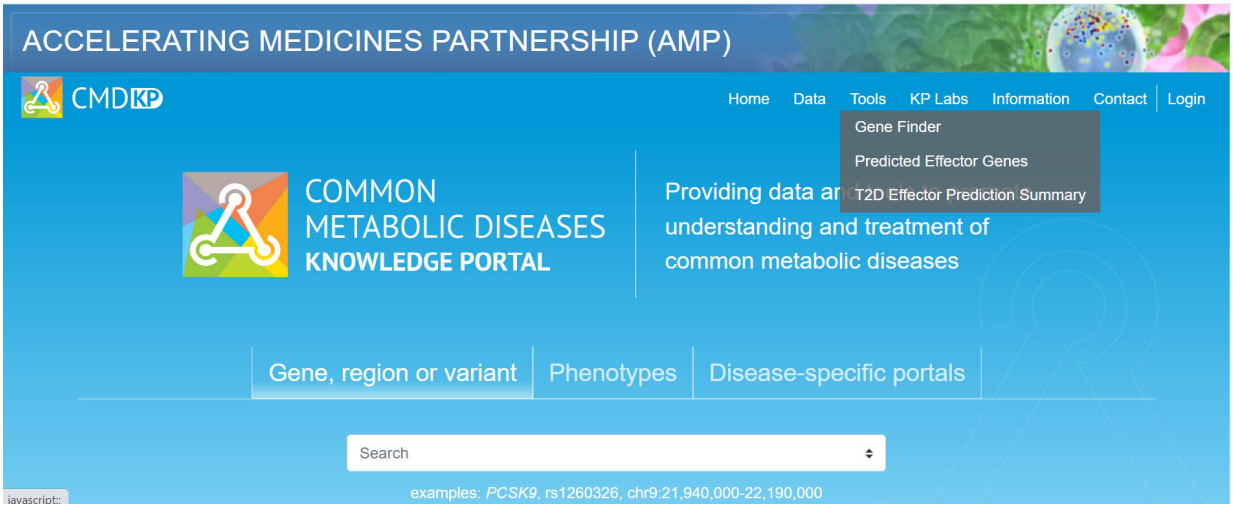
2 Regulatory Evidence

3 Perturbation Evidence

ACCELERATING MEDICINES PARTNERSHIP (AMP)
TYPE 2 DIABETES

78

Where to find the results?



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.

80

Visualize the model & documentation

Curated T2D effector gene predictions

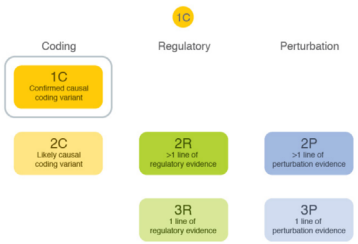
Phenotype

Type 2 diabetes

View data | View research method

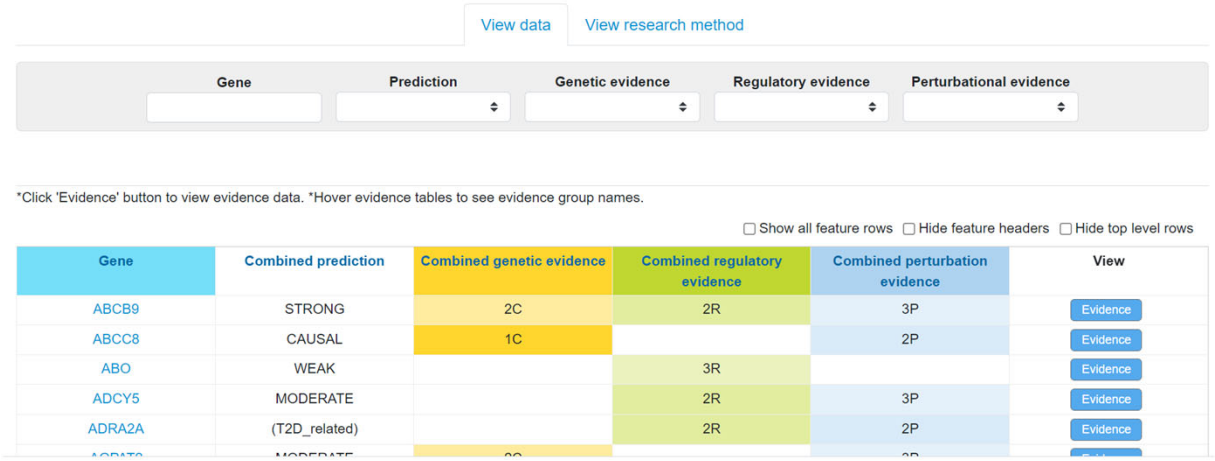
Gene	Prediction	Genetic evidence	Regulatory evidence	Perturbational evidence

Causal



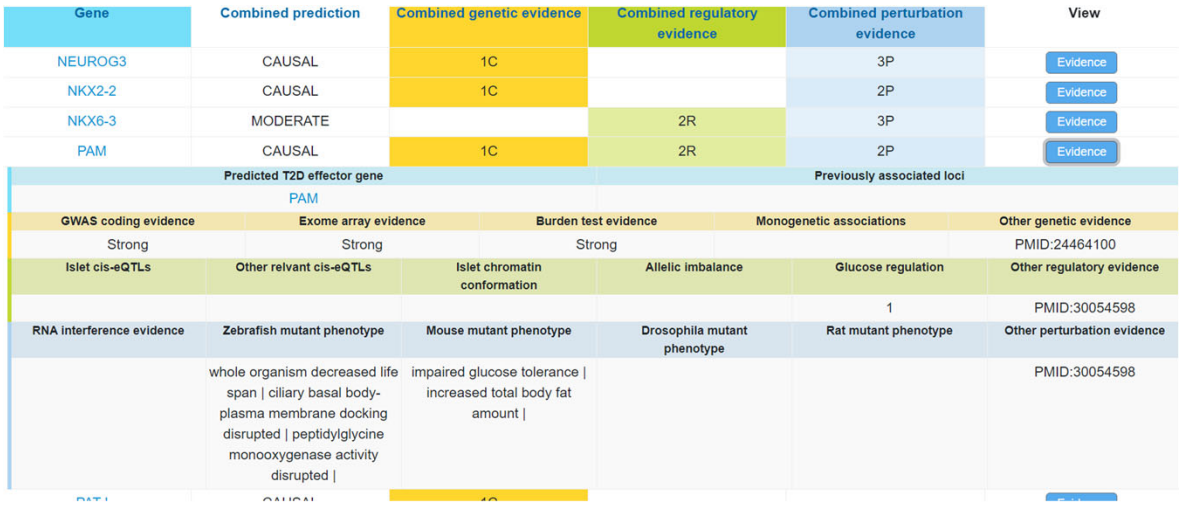
81

Visualize & query the results



82

Visualize the evidence



83

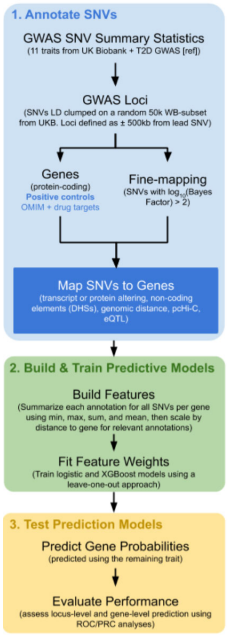
Effector Index Predictions

Brent Richards

Vince Forgetta

An Effector Index to Predict Causal Genes at GWAS Loci

Vincenzo Forgetta*[1], Lai Jiang*[1,8], Nicholas A. Vulpescu*[2], Megan S. Hogan[2], Siyuan Chen[1,8], John A. Morris[1,3,4,10], Stepan Grinek[2], Christian Benner[5], Mark I McCarthy[6‡], Eric Fauman[7], Cecilia MT Greenwood[1,1,8,9,10], Matthew T. Maurano[1,2], J. Brent Richards[1,1,8,10,11,12]



- 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

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-

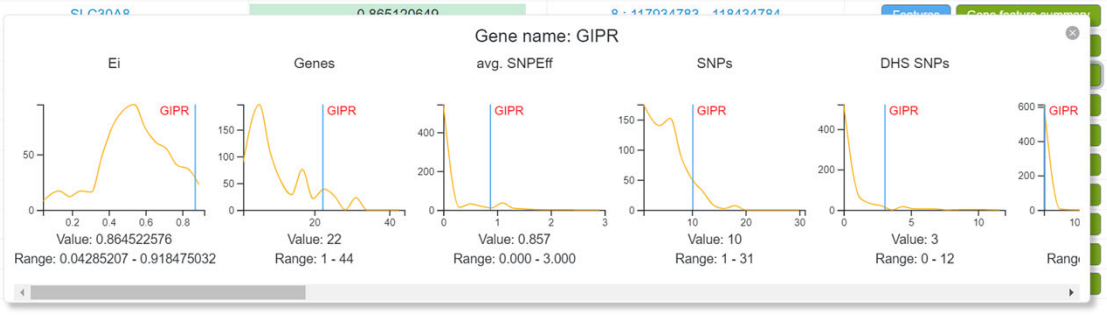
COBLL1	0.860340297	2 : 165258389 - 166012010	Features / Gene feature summary

86

Visualize the evidence

Gene Name	Gene Probability	Locus Location	View
KCNQ1	0.865873158	11 : 1823182 - 3107195	Features / Gene feature summary

Gene name: GIPR



« ‹ 1 2 3 4 … › »

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

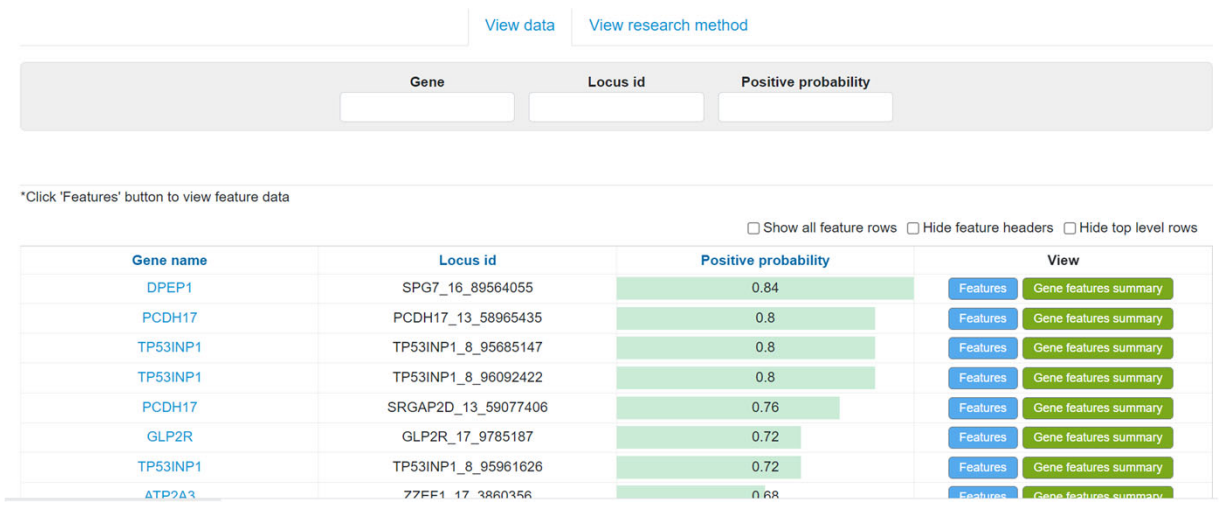
Alisa Manning

Tim Majarian

- 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

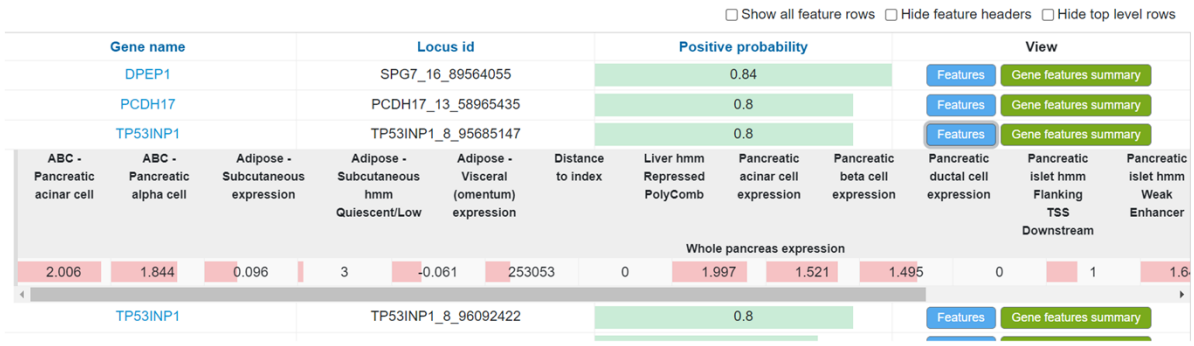
89

Visualize & query the results



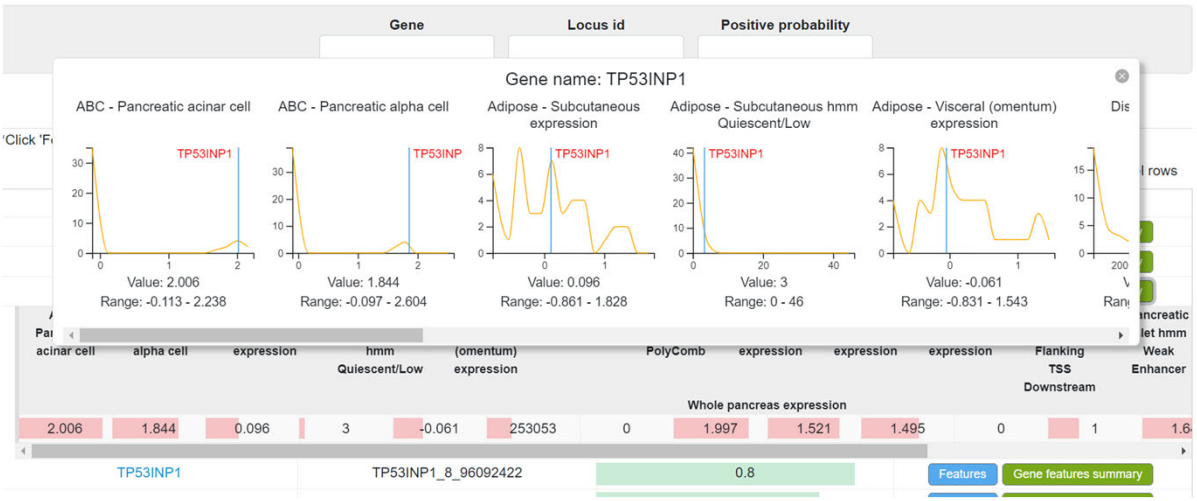
90

Visualize the evidence



91

Visualize the evidence



92

Now what...

Compare, Contrast & Learn

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.

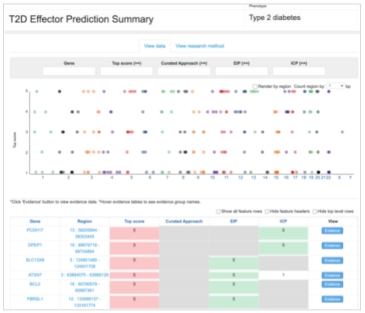
T2D Effector Prediction Summary

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 -	5	5	4		Evidence

96

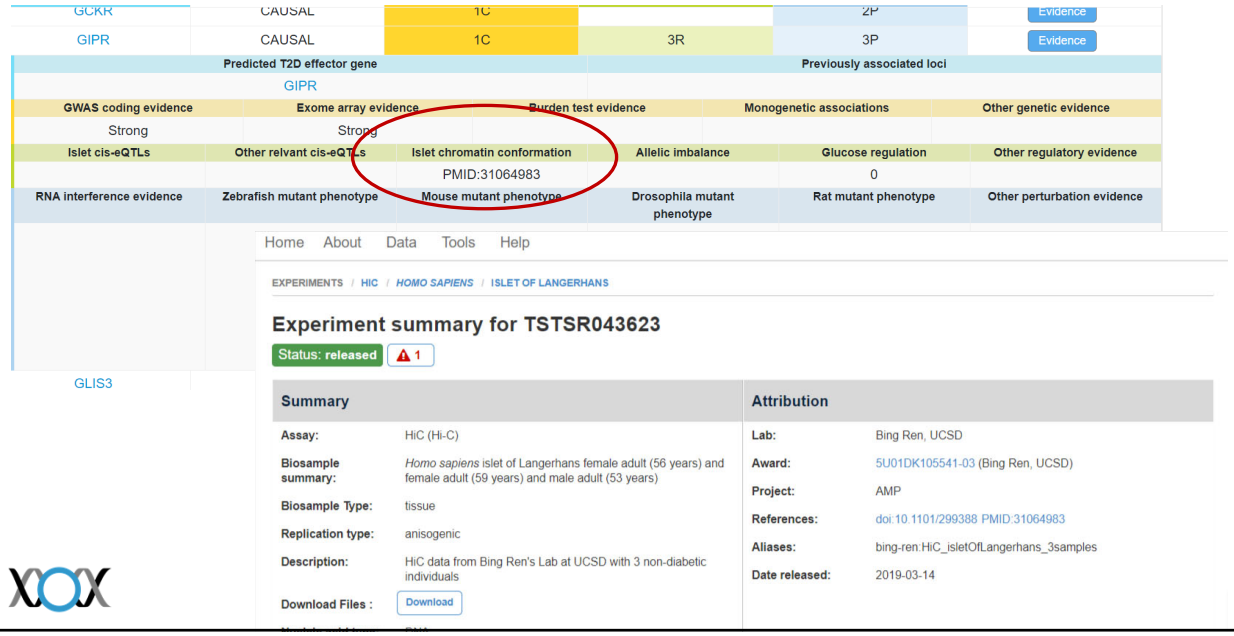
Now what…

HuGeAMP
HUMAN GENETICS AMPLIFIER

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

97

Intregating & making evidence accessible



Current supporting data for the Curated T2D Effector Gene list

Genetic evidence- well captured

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

Regulatory evidence- integration established, limited

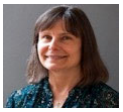
- 11 genome-wide studies: 7 loaded in DGA, in progress for 4
- *Literature evidence from 13 papers for 15 genes*

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

Fully populating the current Curated Effector Gene list


Maria Costanzo

Genetic evidence

- Monogenic associations from OMIM- **load the results to the portal**
- Literature evidence from 11 papers/genes- **load the results to the portal**

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

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

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

CARDIoGRAM Curated Effector Genes

CARDIoGRAMplusC4D (2021) effector gene list

Phenotype: 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,CCDC159,TMEM205	known	known	LDLR	6	LDLR	certain	closest gene; rare coding variants; PoPS; mouse phenotype	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,SMARCD3,ASIC3,AB	known	known	NOS3	6	NOS3	certain	closest gene; PoPS; eQTL; mouse phenotype	Evidence

Number of enriched tissues	Strongest tissue
1	Adrel

95% credible set	Best SNP PPA	Rare variant association	Monogenic disorder	Missense coding variant
10	0.34	LDLR	LDLR, LDLR-AS1	

Top GTEx eQTL gene	Other GTEx eQTL genes	Top STARNET eQTL gene	Other STARNET eQTL genes
LDLR (BLD)	-	-	-

Mouse phenotype	UKBB PheWAS Diseases	UKBB PheWAS Continuous traits	PhenoScanner (non-UKBB)
LDLR		Apolipoprotein_B Cholesterol LDL_direct Triglycerides	

Krishna Aragam

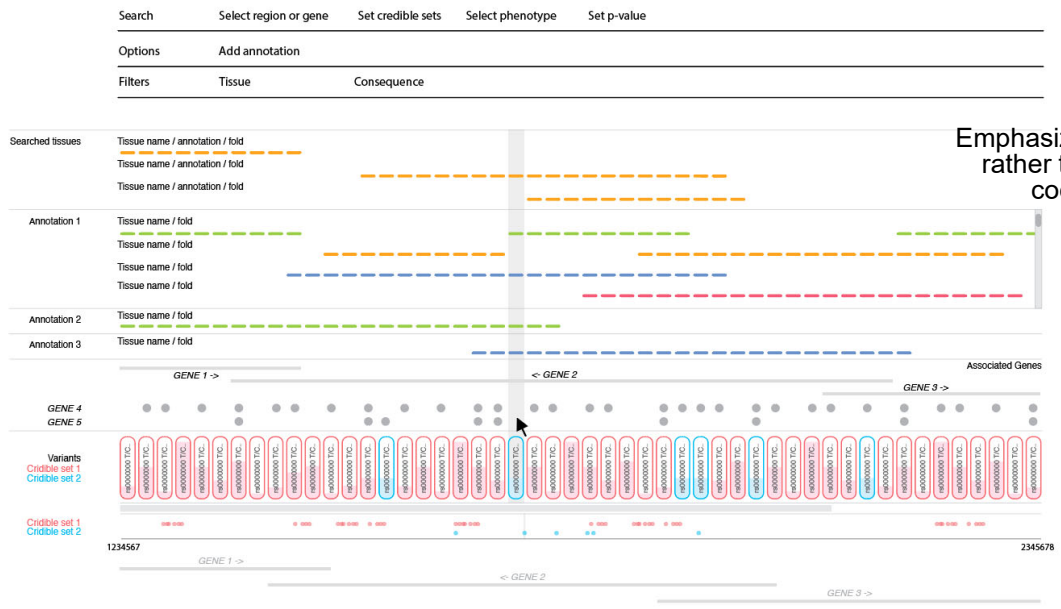
CARDIoGRAMplusC4D

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

Additional views within the portal



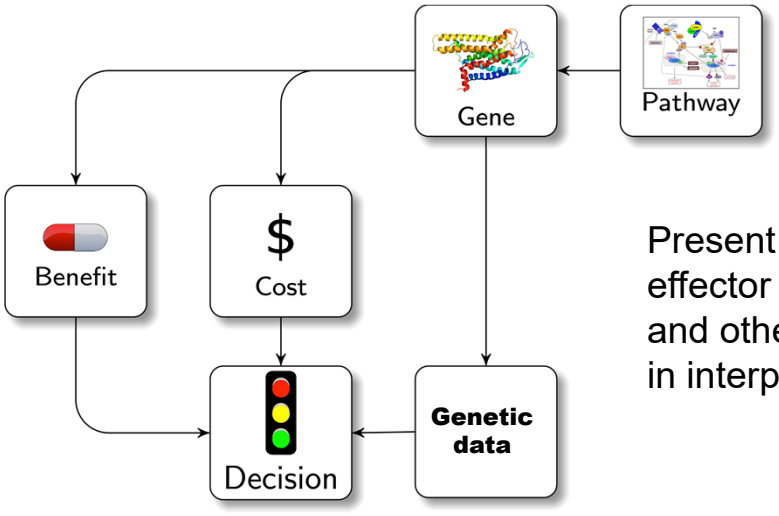
Emphasize relationships rather than genomic coordinates

Advancing prioritizations tools

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

106

Onward…



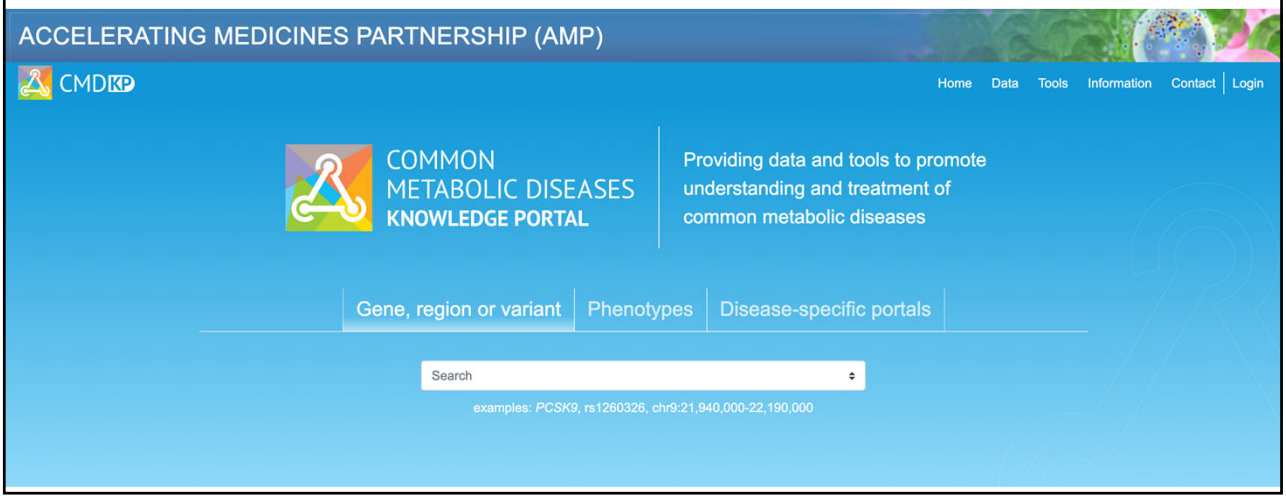
Present integrated views of effector gene predictions, and other data in the portal, in interpretable ways

107

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

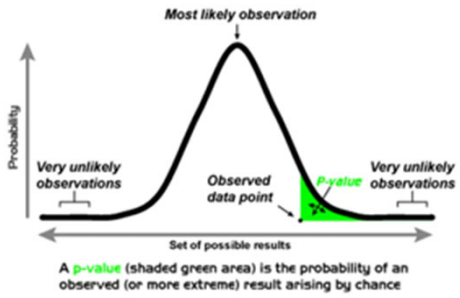
- The T2D Knowledge portal allows users to interrogate a gene across hundreds of traits through a simple web interface

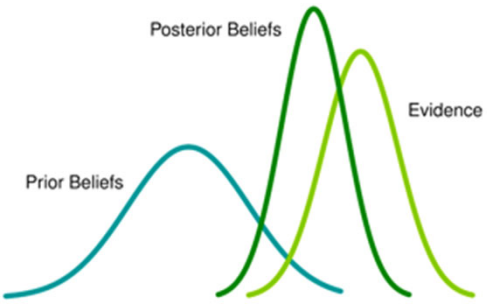
hugeamp.org



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)





Gap 3: what prior should we use?

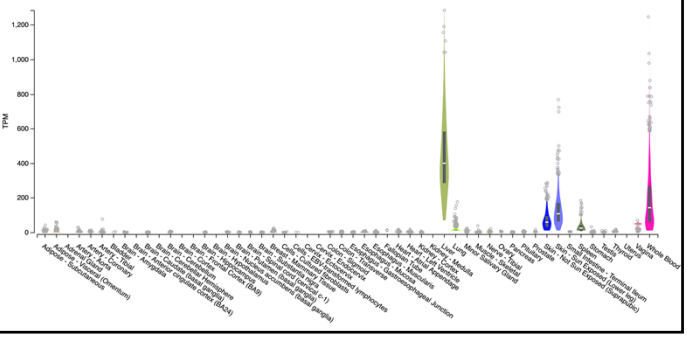
- Most researchers don't have intuition for **quantitative** priors
- But we use **qualitative** priors all the time
- Are you more likely to believe in a T2D-susceptibility gene if…

…it is specifically expressed in liver?

…a mouse knockout has dysglycemia?

Mouse Phenotypes

Availability	Mouse Genotype	decreased insulin secretion	increased fasting circulating gl	hyperglycemia	decreased circulating insulin le	impaired glucose tolerance
Find Mice	Madd^{tm1Bpra}/Madd^{tm1Bpra} Tg(Ins2-cre/ERT)1Dam/0 (conditional)					



110

Membership in a gene set affects a gene's prior

218 genotypes with 218 annotations displayed of selected term and subterms

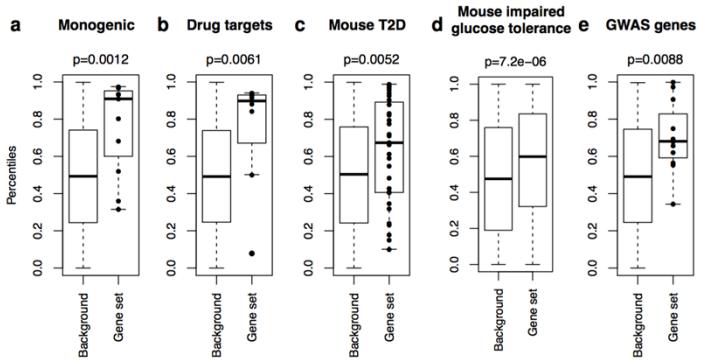
Searched Term: decreased insulin secretion

Allelic Composition (Genetic Background)	Annotated Term	Reference
Abcc8$^{tm1.1Fmas}$/Abcc8$^{tm1.1Fmas}$ (B6.129S2(Cg)-Abcc8$^{tm1.1Fmas}$)	decreased insulin secretion	J:208932
Abcc8$^{tm1.1Mgn}$/Abcc8$^{tm1.1Mgn}$ (involves: 129X1/SvJ * C57BL/6)	decreased insulin secretion	J:79352
Abcc8^{tm1Jbrv}/Abcc8^{tm1Jbrv} (involves: 129X1/SvJ)	decreased insulin secretion	J:179577
Abcc8^{tm1Jbrv}/Abcc8^{tm1Jbrv} Sstr5^{tm1Fcb}/Sstr5^{tm1Fcb} (involves: 129X1/SvJ)	decreased insulin secretion	J:179577
Abhd6$^{tm1a(EUCOMM)Hmgu}$/Abhd6$^{tm1a(EUCOMM)Hmgu}$ (C57BL/6N-Abhd6$^{tm1a(EUCOMM)Hmgu}$)	decreased insulin secretion	J:234876
Ace^{tm3Unc}/Ace^{+} (involves: 129P2/OlaHsd * C57BL/6J)	decreased insulin secretion	J:128859
Adipoq^{tm1Ish}/Adipoq^{tm1Ish} (involves: 129S7/SvEvBrd * C57BL/6J)	decreased insulin secretion	J:199264
Afmid$^{tm1b(EUCOMM)Wtsi}$/Afmid$^{tm1b(EUCOMM)Wtsi}$ (C57BL/6N-Afmid$^{tm1b(EUCOMM)Wtsi}$/H)	decreased insulin secretion	J:226537
Akap5$^{tm1.1Jsco}$/Akap5$^{tm1.1Jsco}$ (involves: 129S * C57BL/6)	decreased insulin secretion	J:190017
Akap5^{tm1Mdag}/Akap5^{tm1Mdag}		

111

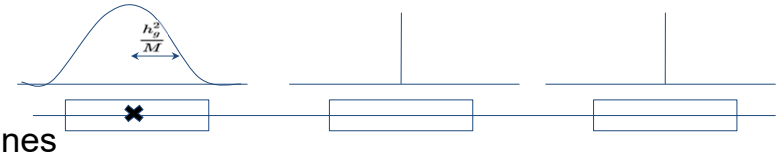
The degree to which should be measurable as a gene set enrichment

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



and/or

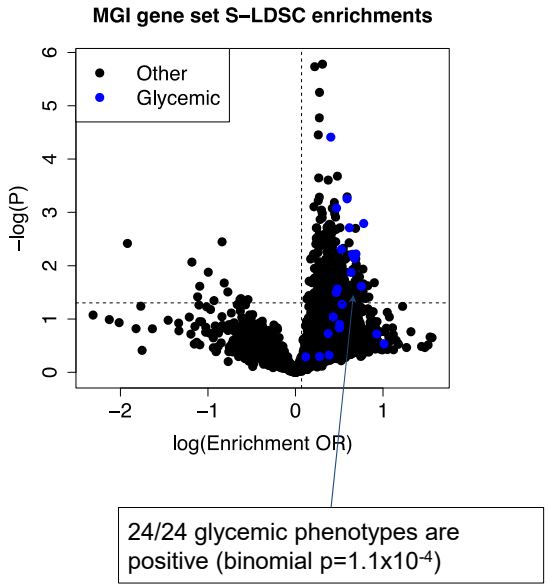
  - Excess heritability explained by **common variants** nearby the genes



112

Application to “mouse knockout” gene sets

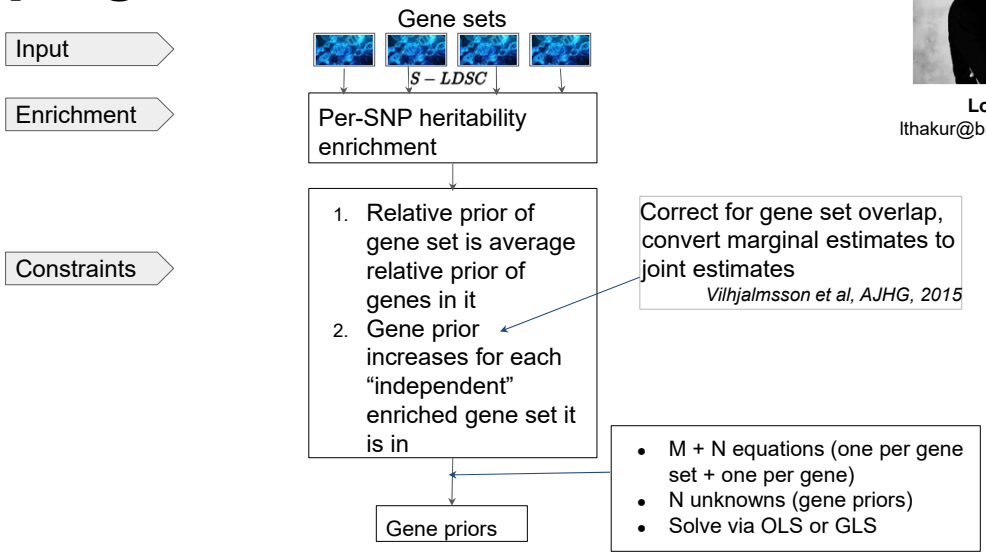
- Calculate heritability enrichments (via Stratified LD-score regression) for each of ~3000 mouse knockout gene sets
- Enrichment odds-ratio roughly corresponds to a “relative prior”



113

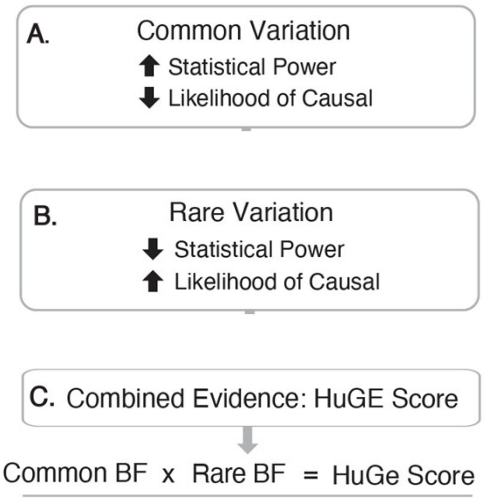
How to combine relative priors from multiple gene sets

Lokendra Thakur
lthakur@broadinstitute.org

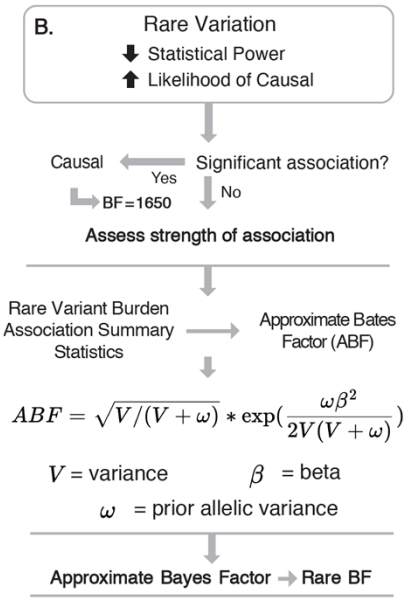


Gap 4: how does genetic data update our prior?

Peter Dornbos
pdornbos@broadinstitute.org



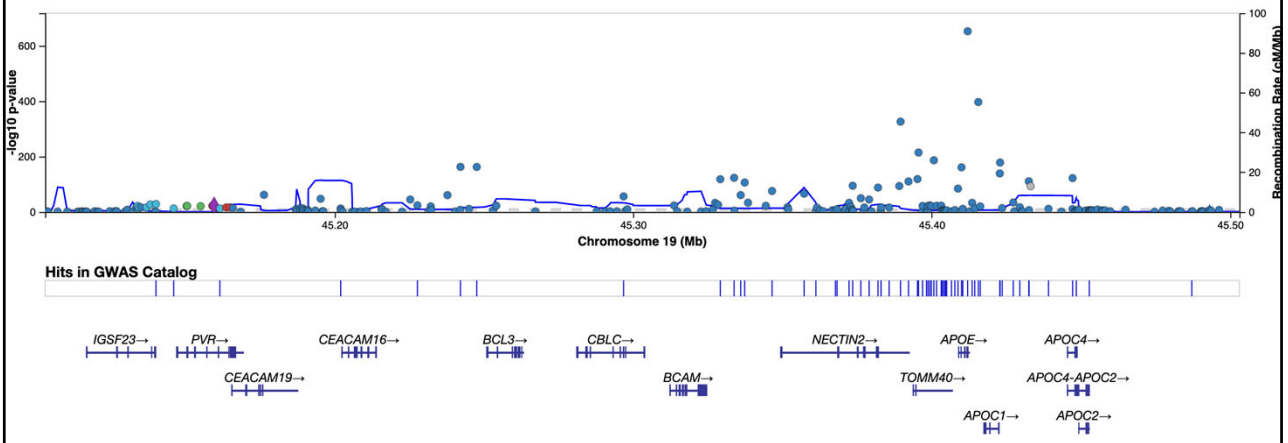
Rare variation



116

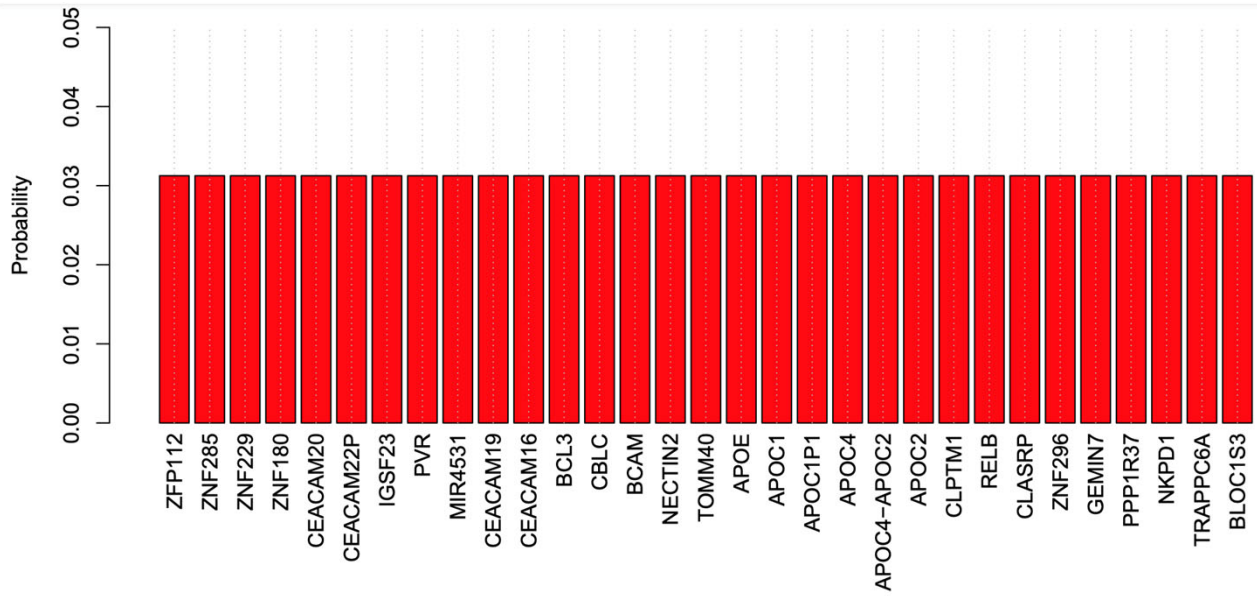
Common variation

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



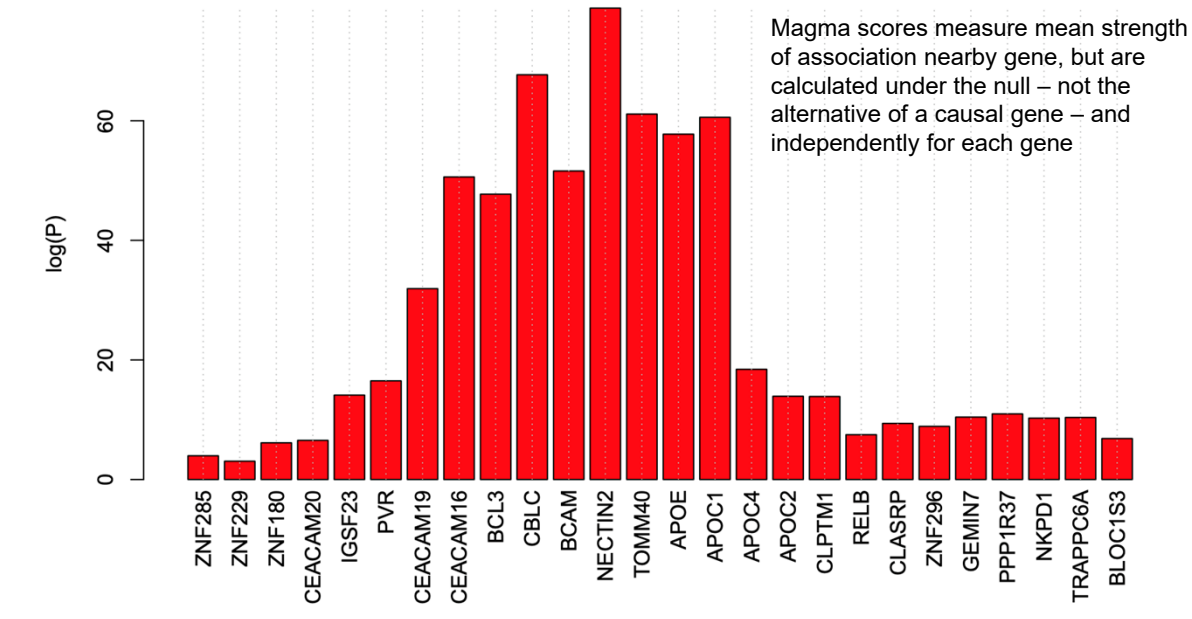
117

Naïve model: uniform likelihood



118

Less naïve model: MAGMA



Magma scores measure mean strength of association nearby gene, but are calculated under the null – not the alternative of a causal gene – and independently for each gene

119

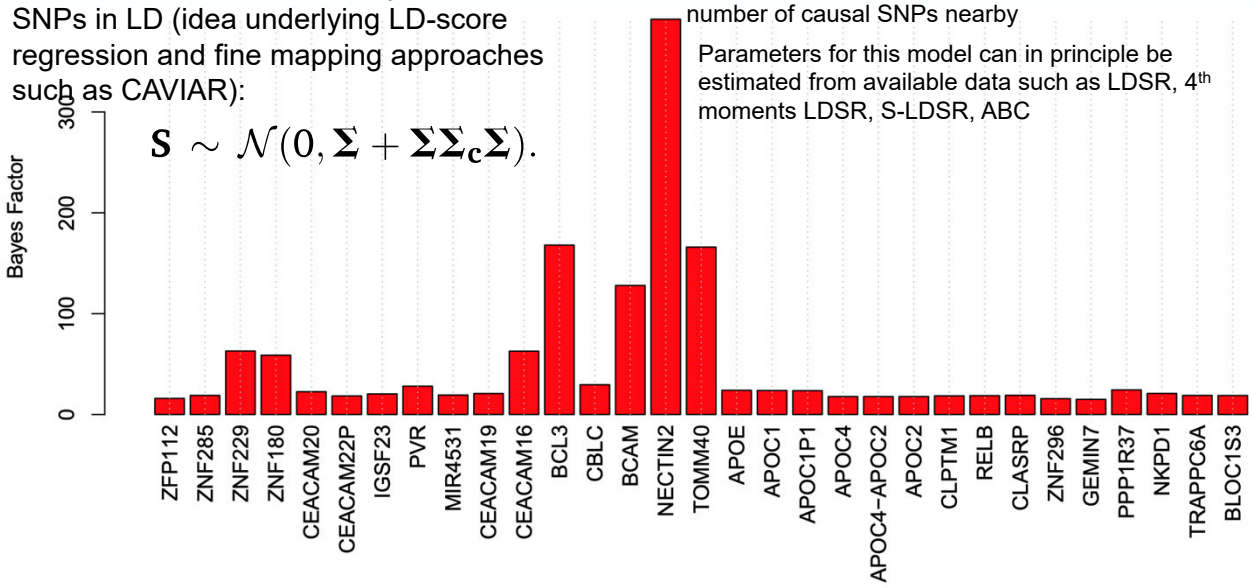
Under development: a Bayesian model

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

$$\mathbf{S} \sim \mathcal{N}(0, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{\Sigma}_{\mathbf{c}}\boldsymbol{\Sigma}).$$

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



Common variation

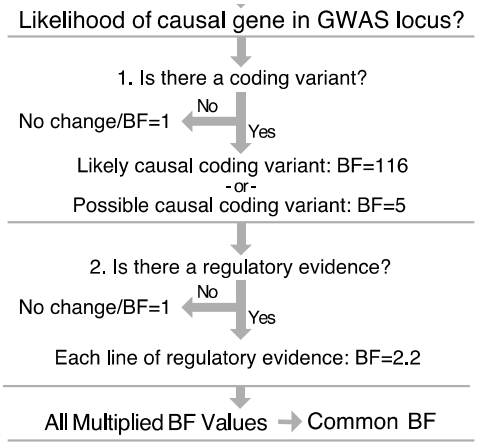
- Second question: is it the effector gene?

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

Converting effector gene predictions to probabilities

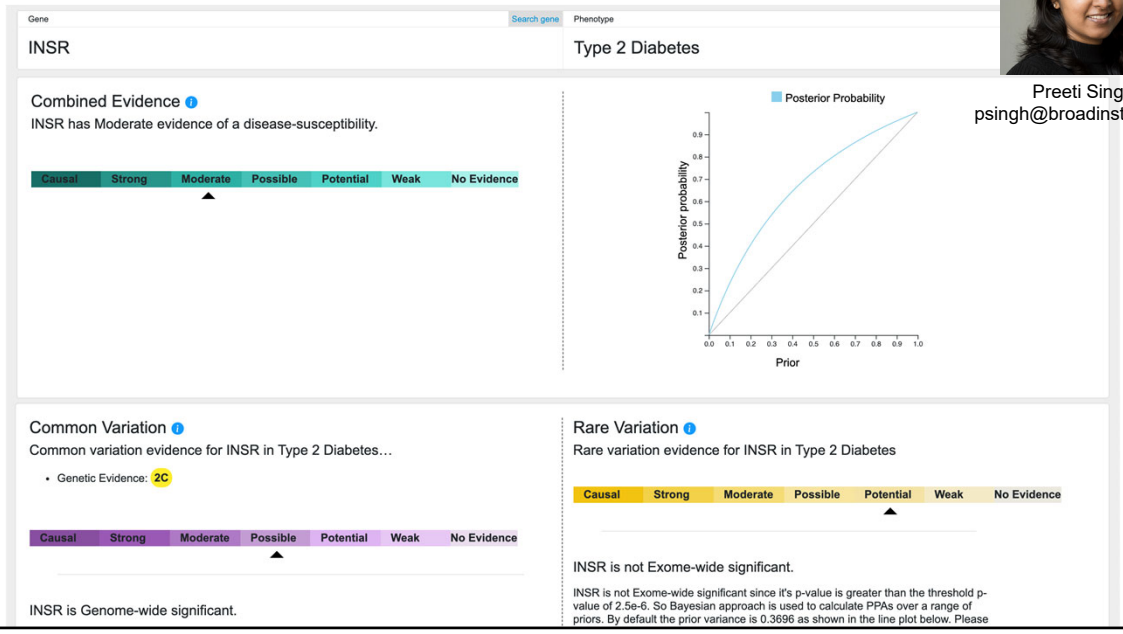
- Some methods directly output probabilities
- Others are qualitative – but, with some transparent assumptions, we can convert them to probabilities
- For the curated list, assuming…
  - "Causal" genes are 95% confident
  - "Strong" genes are 80% confident
  - 2C/2R evidence lines are equivalent
  -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

Then…



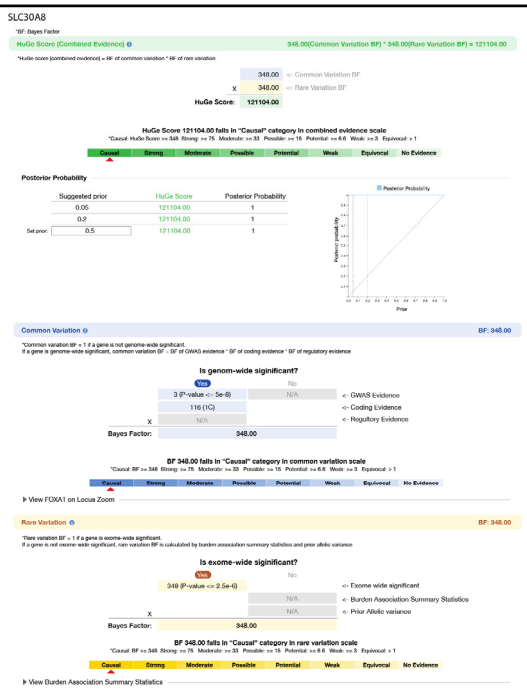
Putting it all together: The HuGE calculator

Preeti Singh
psingh@broadinstitute.org



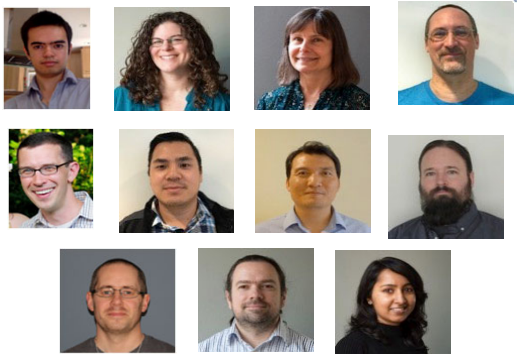
Long term vision

- 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



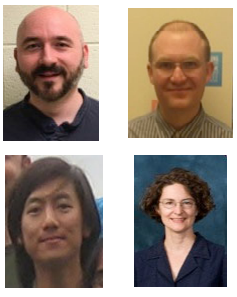
The Multi-Site Team

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

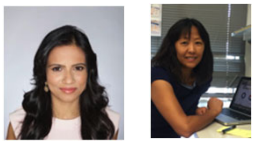


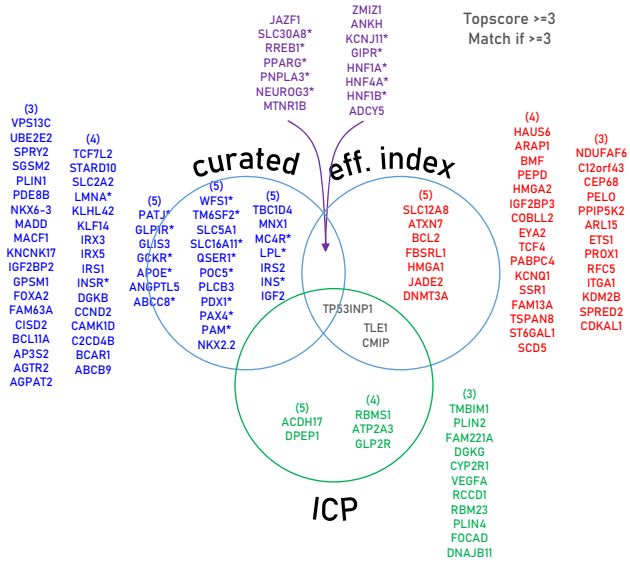
Leadership
Jason Flannick
Noël Burtt
Mike Boehnke
Kyle Gaulton
Thomas Keane
Jose Florez

UM
Ryan Welch
Andy Boughton
Alan Kwong
Laura Scott



UCSD/DGA
Parul Kudtarkar
Ying Sun
Sharvari Narendra





	curated	EI	ICP
1: 40Mb	MACF1	PABPC4	
2: 65Mb		CEP68 SPRED2	
3: 185Mb	IGF2BP2	ST6GAL1	DGKG DNAJB11
5: 102Mb	PAM	PPIP5K2	
6: 7Mb	RREB1	RREB1 SSR	
6: 39Mb	GLP1R KCNK17		
7:23Mb		IGF2BP3	FAM221A
8: 95Mb		TP53INP1 NDUFAF6	TP53INP1
9:19Mb		HAUS6	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	

PANEL DISCUSSION

FNIH
Foundation for the National Institutes of Health