臺灣大型基因晶片 資料比較

Chien-Hsiun Chen, Ph.D.

2024 NCGM服務說明會

NCGM, IBMS, Academia Sinica

2024-01-11

Taiwan Biobank

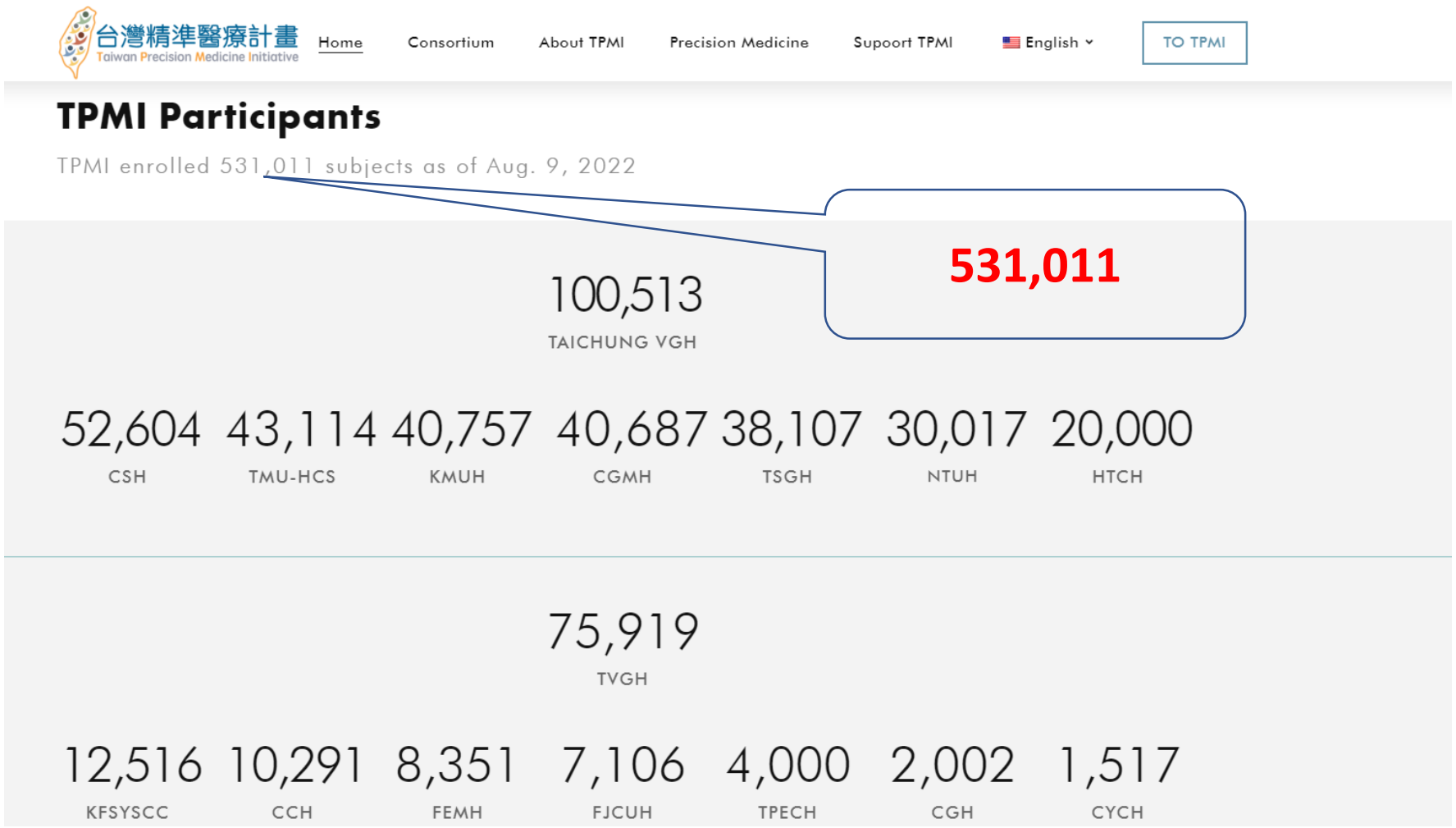
臺灣人體生物資料庫

- 200,000 normal participants
- 100,000 patients (10~15 diseases)

https://www.twbiobank.org.tw/



Taiwan Precision Medicine Initiative



TWB 2.0 Array (TPM 1.0 Array) and TPM 2.0 Array

	TWB 2.0	TPM 2.0
probe SNP	752,921 (686,463 SNPs)	755,191 (743,227 SNPs)
SNPs 挑選之來源	● TWB Array，共104,909個SNP點。 ● 根據TWB Array的使用經驗及Taiwan Biobank 1000名參與者次世代定序的結果，設計446K個SNP點。 ● 與臨床顯著相關的SNP點約105K個。 ● Thermo Fisher Scientific歷年替各國Biobank設計的疾病相關標記SNP。 ● 其他與藥物反應、藥物代謝相關，如MHC，PGX等基因上的SNP點。 ● 偵測拷貝數變異（CNV）的SNP點。	● 根據Taiwan Biobank次世代定序及Han Chinese次世代定序的結果，設計596K個SNP點。 ● 與臨床顯著相關的SNP點約124K個。 ● Thermo Fisher Scientific歷年替各國Biobank設計的疾病相關標記SNP。 ● 其他與藥物反應、藥物代謝相關，如MHC，PGX等基因上的SNP點。 ● 偵測拷貝數變異（CNV）的SNP點。
【註】兩款 array plate相同的SNP點約496K個。		

http://ncgm.sinica.edu.tw/affymetrix_tech_01.html

The Data Sets

	Total N	Array	# of Subjects	# of Variants
TWB	147,895	TWB1	27,732	646,973
		TWB2	120,163	686,439
TPMI	487,026	TPM1	165,666	686,439
		TPM2	321,360	725,457

Batch effect

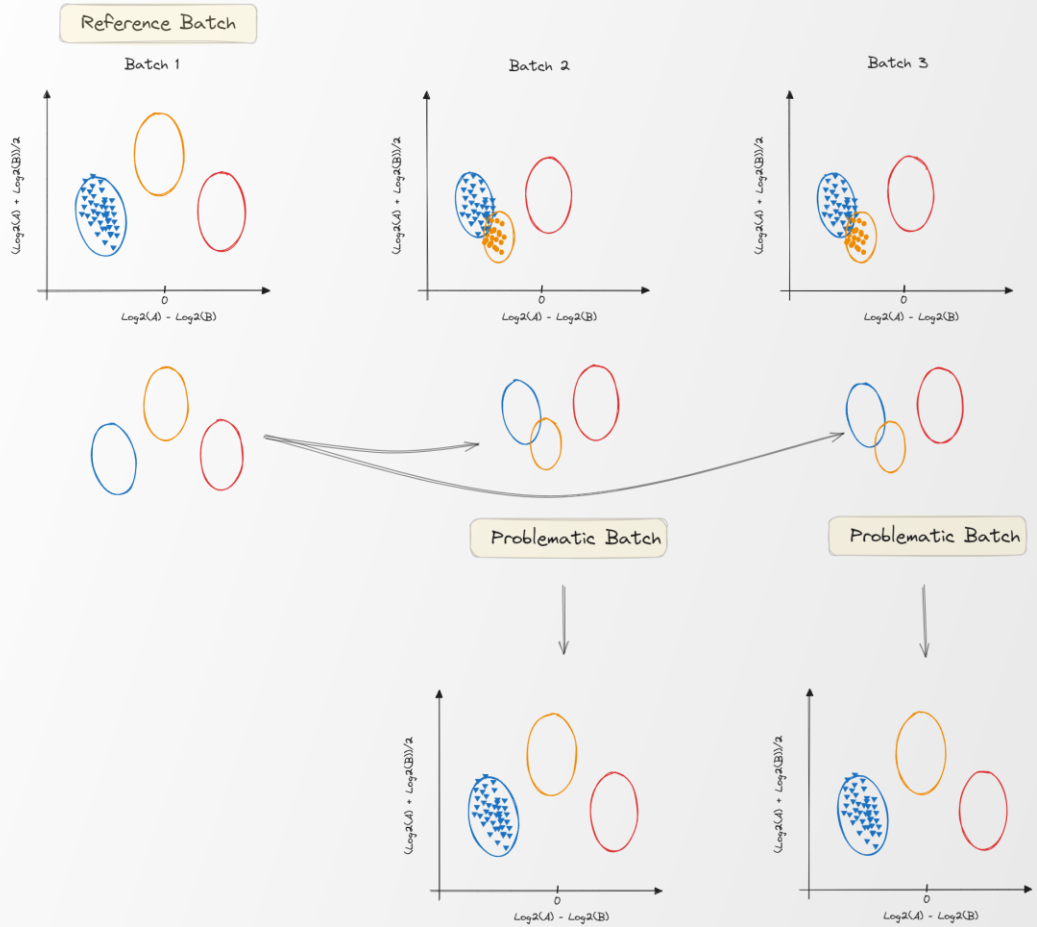
- Different genotyping time
  - Different lots of arrays
  - Different reagents
  - Different operators
- Cases and controls
  - User's samples
  - Using biobanking samples
- Genotype calling
  - Individual plate
  - Pooling plates
- Annotations
- Solution: Using cross-platform validation

CMOD - TPMI Final Callset Tuning Strategy

Step1: Identify the Reference Batch using MAF

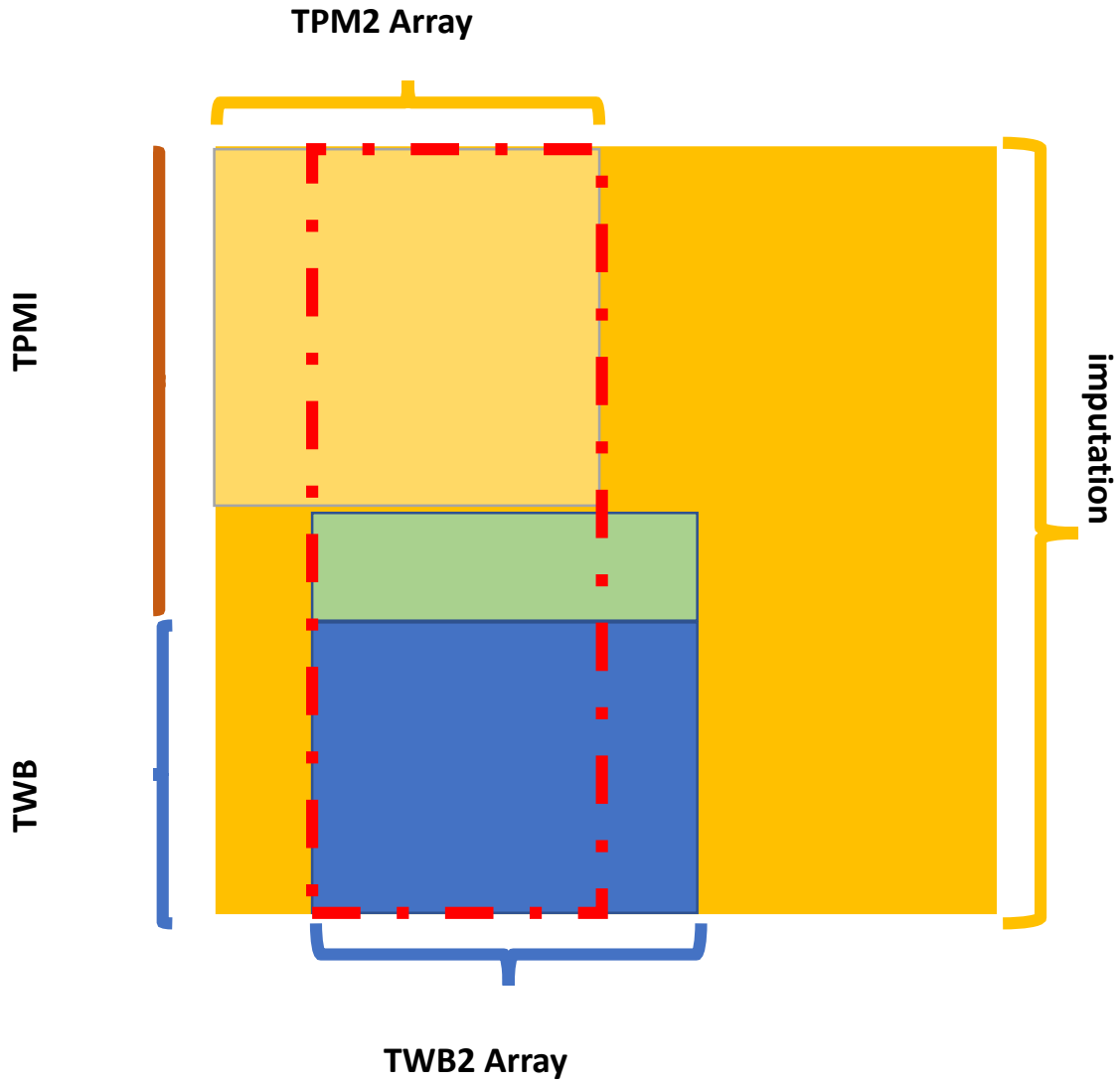
Step2: Find Problematic Batches Using Cosine Similarity of Posteriors

Step3: Multiplate Normalization based on Signals from the Reference Batch



Imputation Steps

1. Genotype QC
2. Merging array with overlapping variants
3. Phasing haplotype by SHAPEIT5
4. Imputation by IMPUTE5
   - Total of 20M variants are imputed
   - Total of 8M variants pass through INFO>0.7 and MAF>0.01

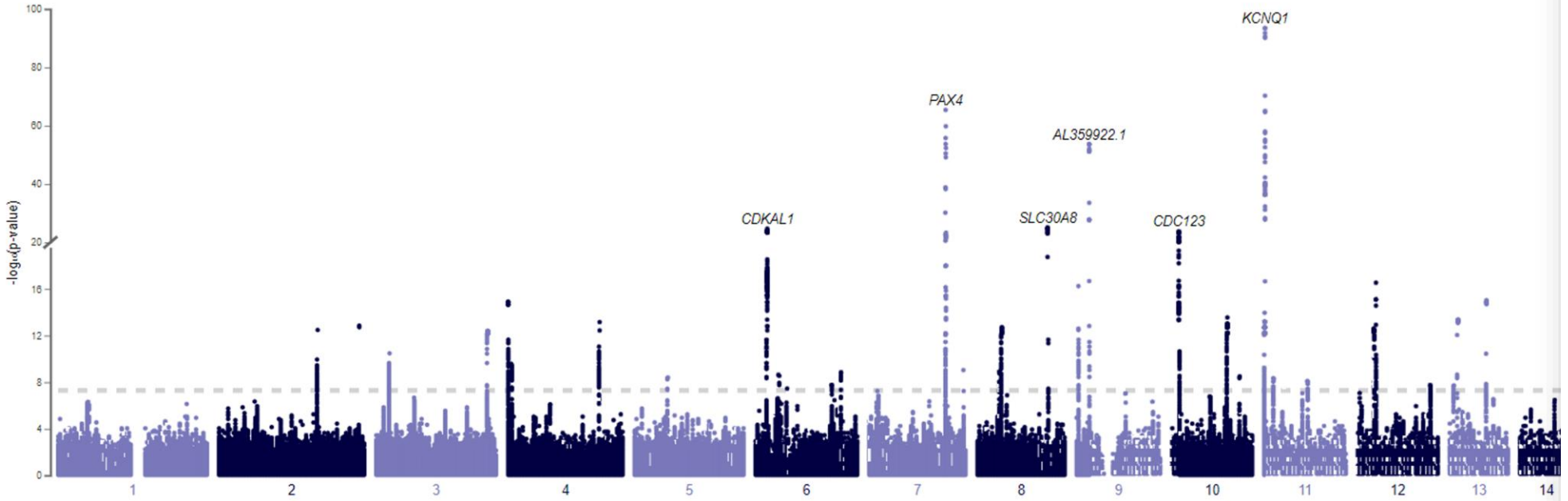


TPMI PheWeb

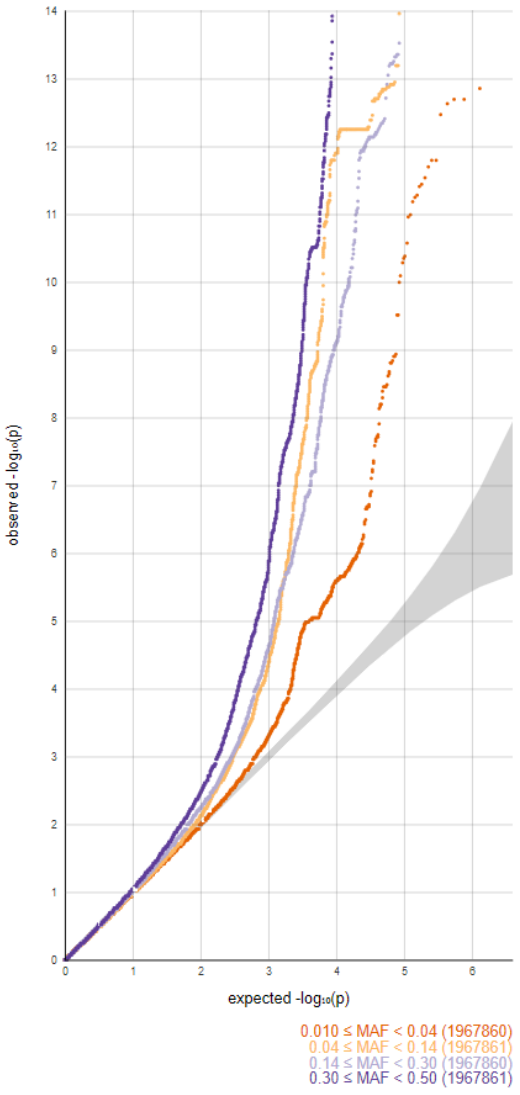
Search for a variant, gene, or phenotype

250.2: Type 2 diabetes

49397 cases, **216833** controls
Category: **endocrine/metabolic**
Manhattan QQ



QQ plot:



Top Loci:

Search... "TCF7L2", "rs1861867", etc.

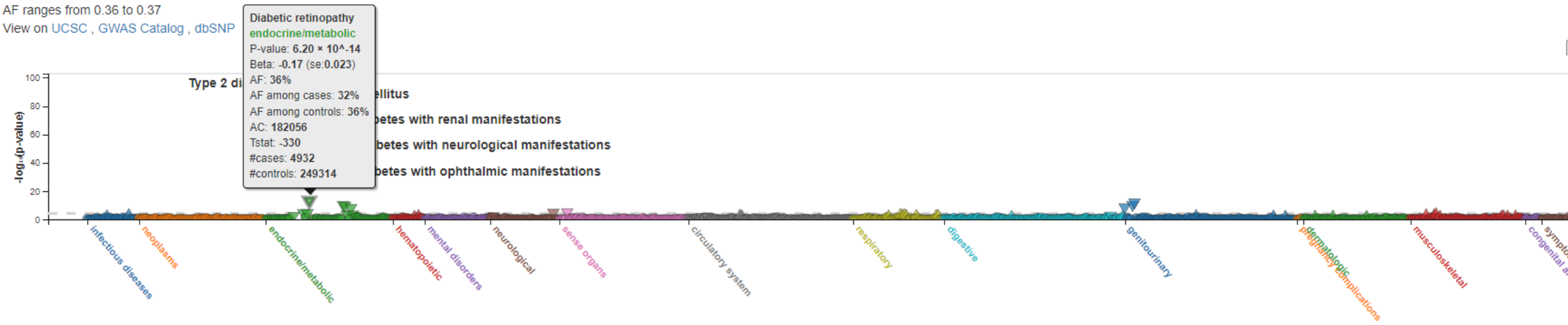
Variant	Nearest Gene(s)	MAF	P-value
11:2,836,003 G / A (rs60808706)	*KCNQ1*	0.36	3.0e-94
7:127,613,496 C / T (rs2233580)	*PAX4*	0.11	3.7e-66
9:22,134,095 T / C (rs10811661)	*AL359922.1*	0.41	2.0e-54

TPMI PheWeb

Search for a variant, gene, or phenotype

11 : 2,836,003 G / A (rs60808706)

Nearest gene: *KCNQ1*
AF ranges from 0.36 to 0.37
View on UCSC , GWAS Catalog , dbSNP



Search... "427.21", "Diabetes", etc.

Category	Phenotype	P-value	Effect Size (se)	Number of samples
endocrine/metabolic	Type 2 diabetes	3.0e-94	-0.19 (0.0092)	49397 / 216833
endocrine/metabolic	Diabetes mellitus	4.1e-83	-0.17 (0.0086)	55703 / 216847
endocrine/metabolic	Type 2 diabetes with renal manifestations	9.1e-56	-0.21 (0.013)	18616 / 216833
endocrine/metabolic	Type 2 diabetes with ophthalmic manifestations	5.4e-46	-0.21 (0.015)	13647 / 216833
endocrine/metabolic	Type 2 diabetes with neurological manifestations	5.8e-46	-0.20 (0.014)	15201 / 216833
endocrine/metabolic	Other disorders of pancreatic internal secretion	7.5e-23	-0.16 (0.016)	10376 / 255880
endocrine/metabolic	Diabetic retinopathy	6.2e-14	-0.17 (0.023)	4932 / 249314

Acknowledgements



