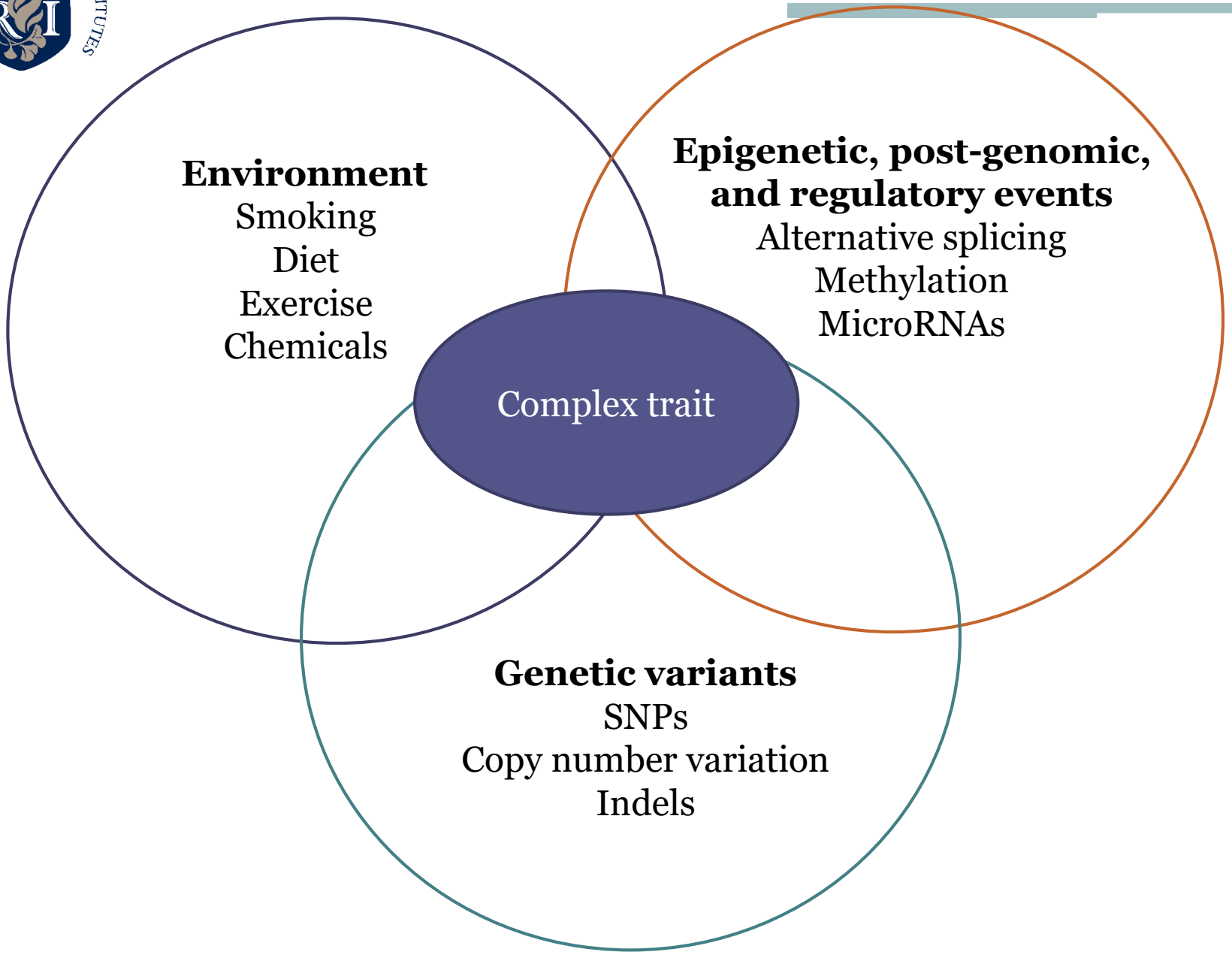
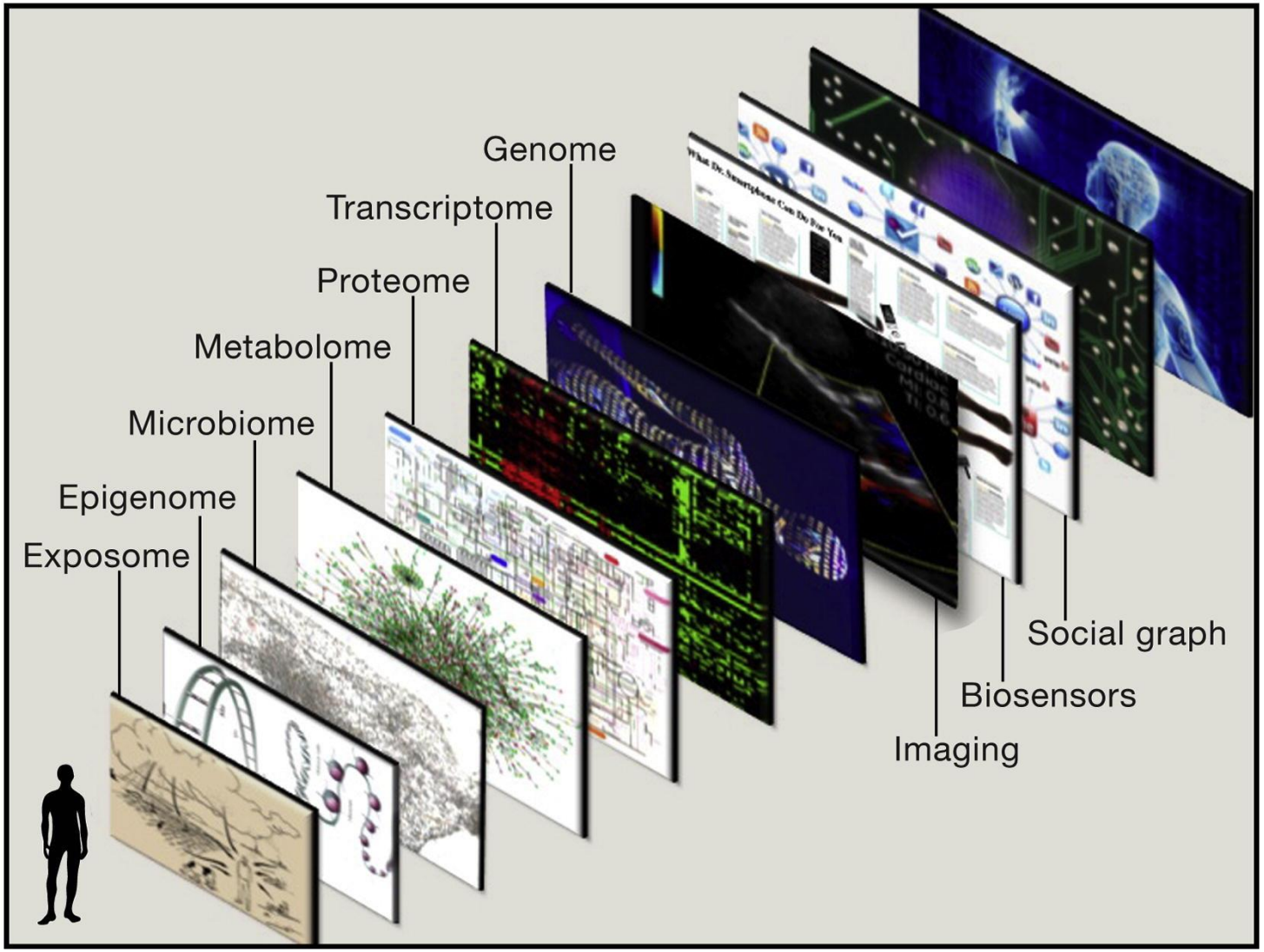
A Large-scale Genetic Association Study for Fasting Glucose from Multiple Cohorts

Ren-Hua Chung
Associate Investigator
Institute of Population Health Sciences
National Health Research Institutes



http://www.ucsf.edu/news/2011/08/10431/major-genetic-study-multiple-sclerosis-reveals-dna-hot-spots-disease

Biomedical big data



Topol. 2014. Cell.

The All of Us research program

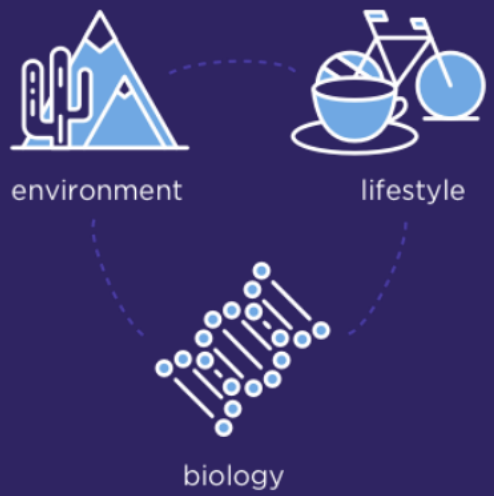
NIH National Institutes of Health
All of Us Research Program

ABOUT FUNDING NEWS, EVENTS, & MEDIA JoinAllofUs.org › Search

We are building a research program of 1,000,000+ people

The mission of the *All of Us* Research Program is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us.

ABOUT THE SCALE & SCOPE



Research focuses on the intersection of 3 factors

臺灣
Taiwan Biobank

The Cumulative
累計收案數

統計至2018年5月31日止(請按此)

社區民眾收案數

98,813

參與個案總數

19,390

完成第一輪追蹤個案總數

醫學中心患者收案數

1,087

參與個案總數

66

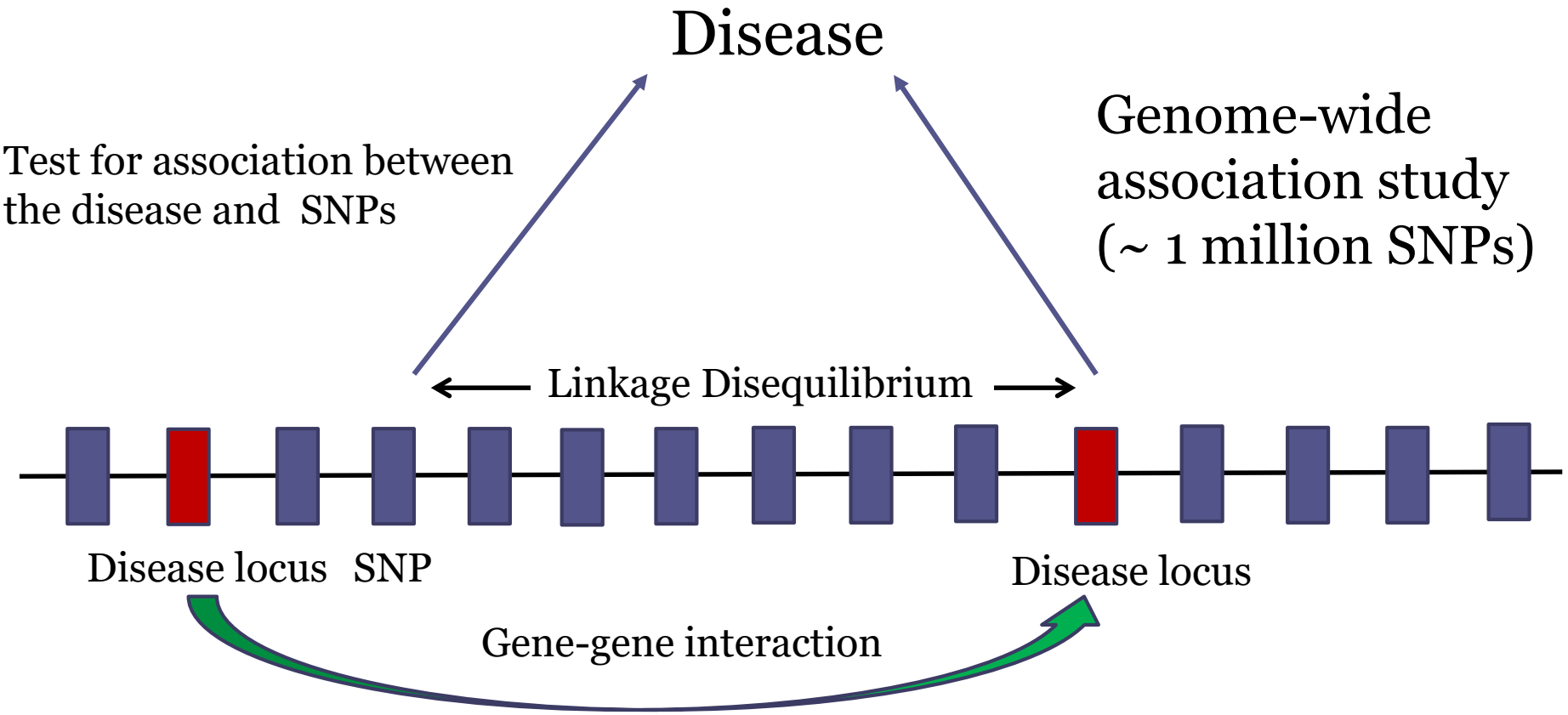
完成第一輪追蹤個案總數

Thermo Fisher: Ion Proton

Illumina: HiSeq 2500

Affymetrix TWB chip

Genetic association analysis

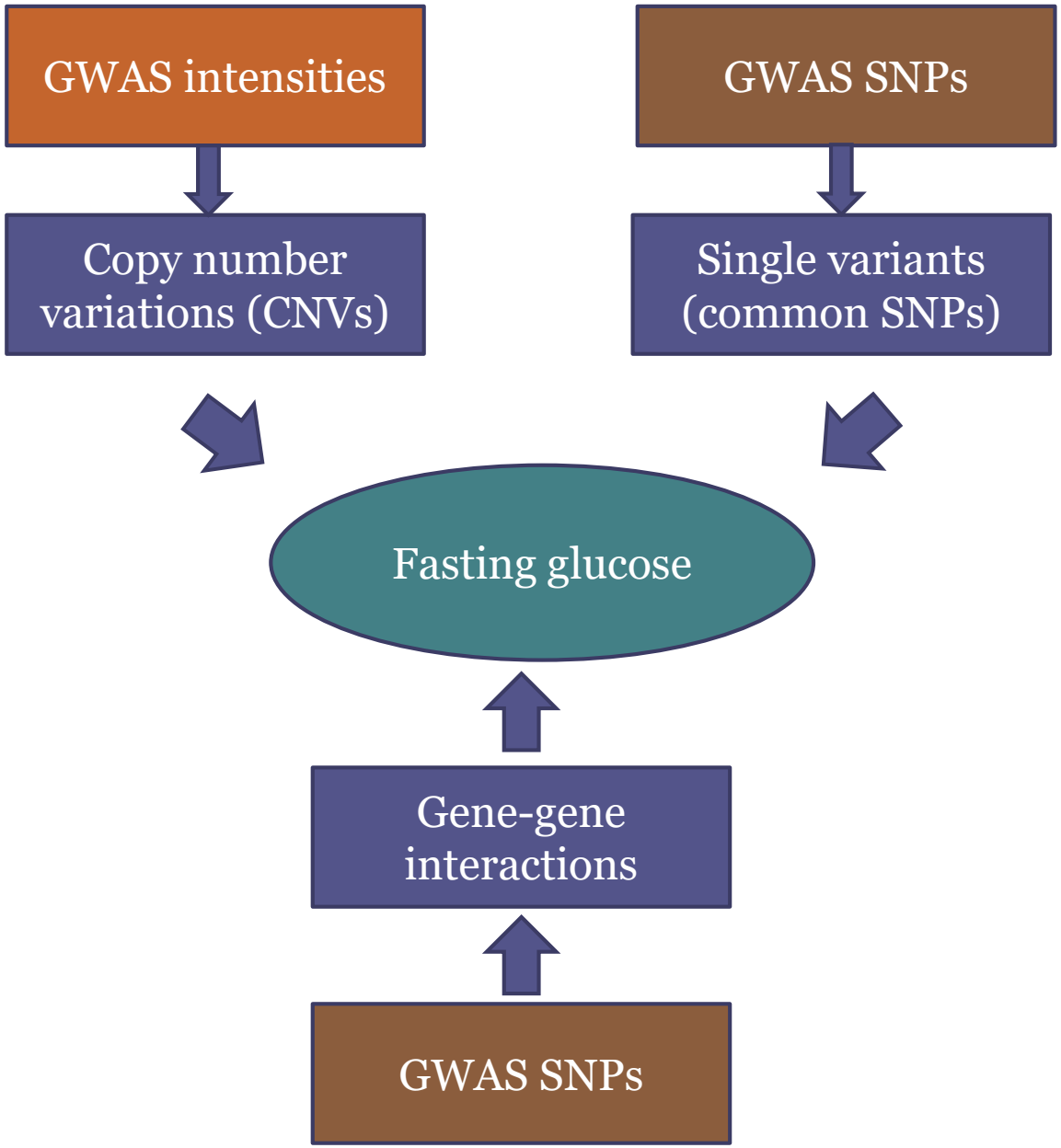


GWAS cohorts

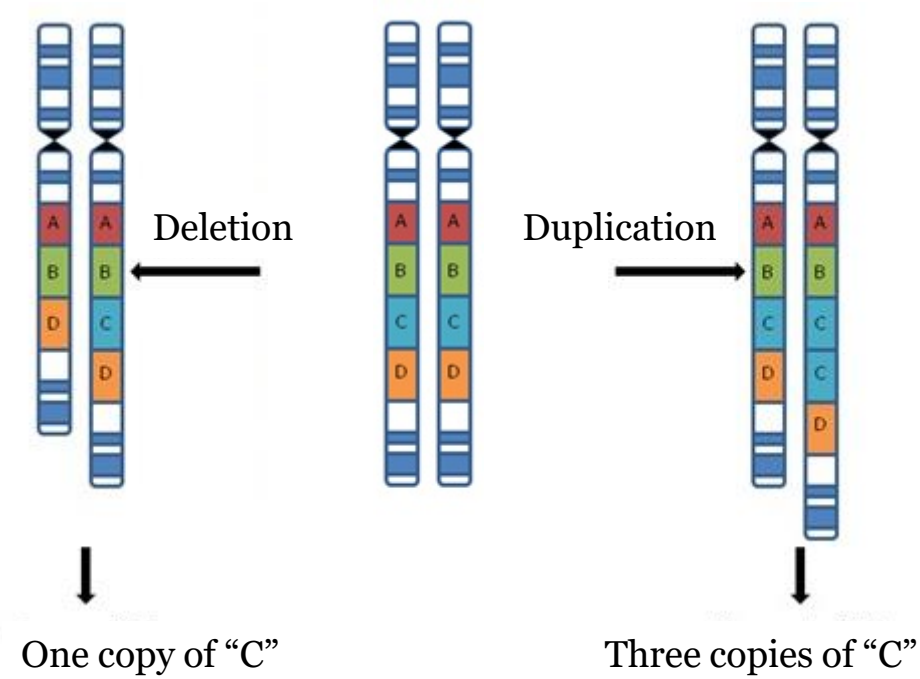
- Stanford Asia-Pacific Program for Hypertension and Insulin Resistance (SAPPHIRe) family study (Wu et al., 2002)
  - Sample size: 440. Number of SNPs: 666,732
- The Healthy Aging Longitudinal Study in Taiwan (HALST) (Hsu et al., 2017)
  - Sample size: 1,883. Number of SNPs: 91,305
- Taiwan Biobank (TWB) (Fan et al. 2008)
  - Sample size: 12,435. Number of SNPs: 592,131

Fasting glucose

- Measured after an 8-10 hour overnight fast.
- Impaired fasting glucose is considered as early stages of diabetes.
- Heritability in families with hypertension (Freedman et al., 2005): 52%
- Proportion of heritability explained by GWAS SNPs (Vattikuti et al., 2012): 30%


GWAS intensities
GWAS SNPs
Copy number variations (CNVs)
Single variants (common SNPs)
Fasting glucose
Gene-gene interactions
GWAS SNPs

Copy number variations (CNVs)



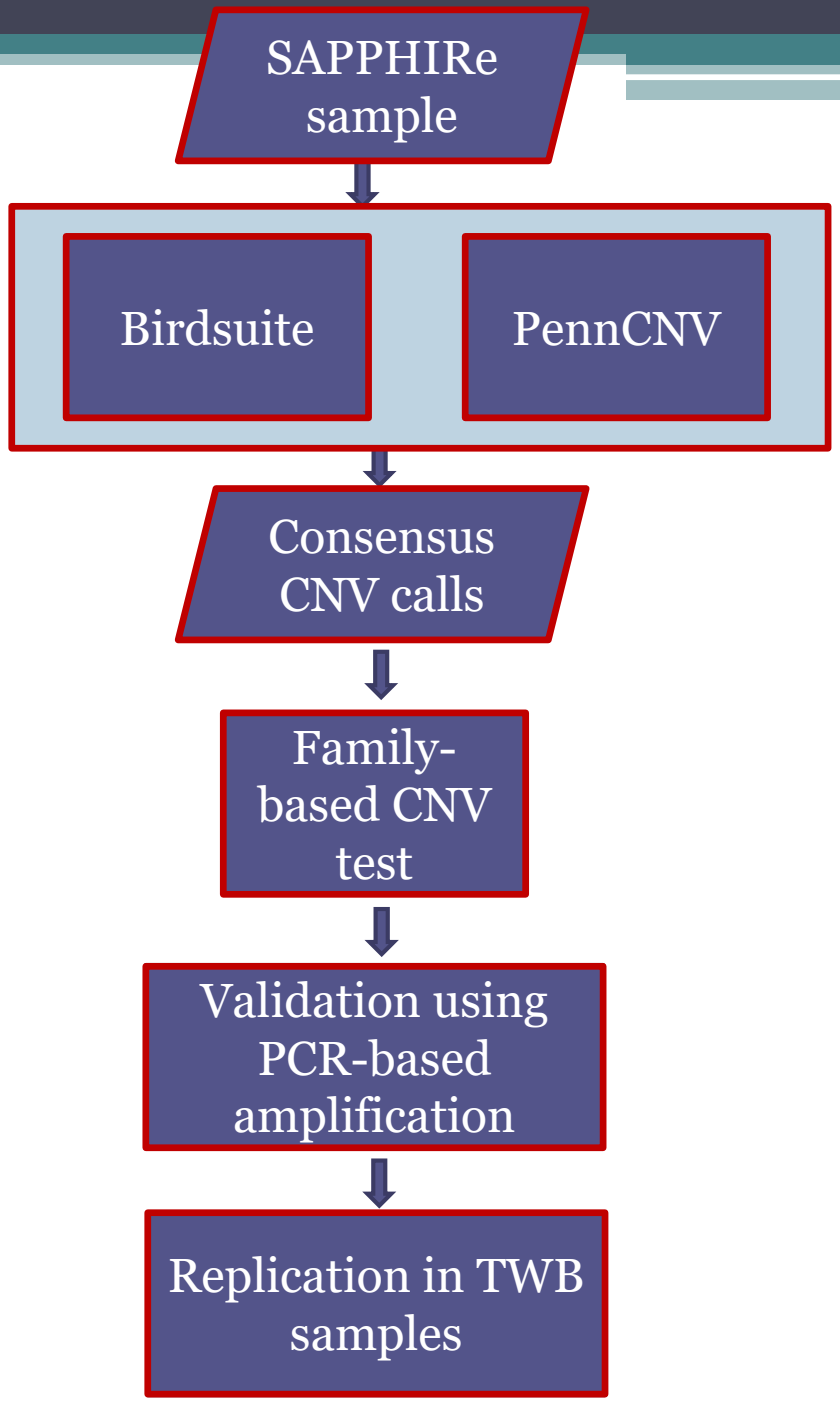
http://readingroom.mindspec.org/?page_id=8221

RESEARCH ARTICLE Open Access

Genome-wide copy number variation analysis identified deletions in SFMBT1 associated with fasting plasma glucose in a Han Chinese population

Ren-Hua Chung[1], Yen-Feng Chiu[1], Yi-Jen Hung[2], Wen-Jane Lee[3,4], Kwan-Dun Wu[5], Hui-Ling Chen[1], Ming-Wei Lin[6], Yii-Der I. Chen[7], Thomas Quertermous[8] and Chao A. Hsiung[1*]

Results: We conducted a genome-wide CNV association analysis for fasting plasma glucose (FPG) and fasting plasma insulin (FPI) using a family-based genome-wide association study sample from a Han Chinese population in Taiwan. A family-based CNV association test was developed in this study to identify common CNVs (i.e., CNVs with frequencies ≥ 5%), and a generalized estimating equation approach was used to test the associations between the traits and counts of global rare CNVs (i.e., CNVs with frequencies <5%). We found a significant genome-wide association for common deletions with a frequency of 5.2% in the Scm-like with four mbt domains 1 (*SFMBT1*) gene with FPG (association *p*-value = 2×10^{-4} and an adjusted *p*-value = 0.0478 for multiple testing). No significant association was observed between global rare CNVs and FPG or FPI. The deletions in 20 individuals with DNA samples available were successfully validated using PCR-based amplification. The association of the deletions in *SFMBT1* with FPG was further evaluated using an independent population-based replication sample obtained from the Taiwan Biobank. An association *p*-value of 0.065, which was close to the significance level of 0.05, for FPG was obtained by testing 9 individuals with CNVs in the *SFMBT1* gene region and 11,692 individuals with normal copies in the replication cohort.


SAPPHIRe sample
Birdsuite
PennCNV
CNV calling
Consensus CNV calls
Family-based CNV test
Validation using PCR-based amplification
Replication in TWB samples

CNV analysis results in *SFMBT1*

SAPPHIRe			
CNV	**Mean glucose level**	**Lower.CL**	**Upper.CL**
Deletion	89.4	83.7	95.1
Normal	93.7	89.0	98.4
P-value: 0.0002. Adjusted p-value: 0.0478.			

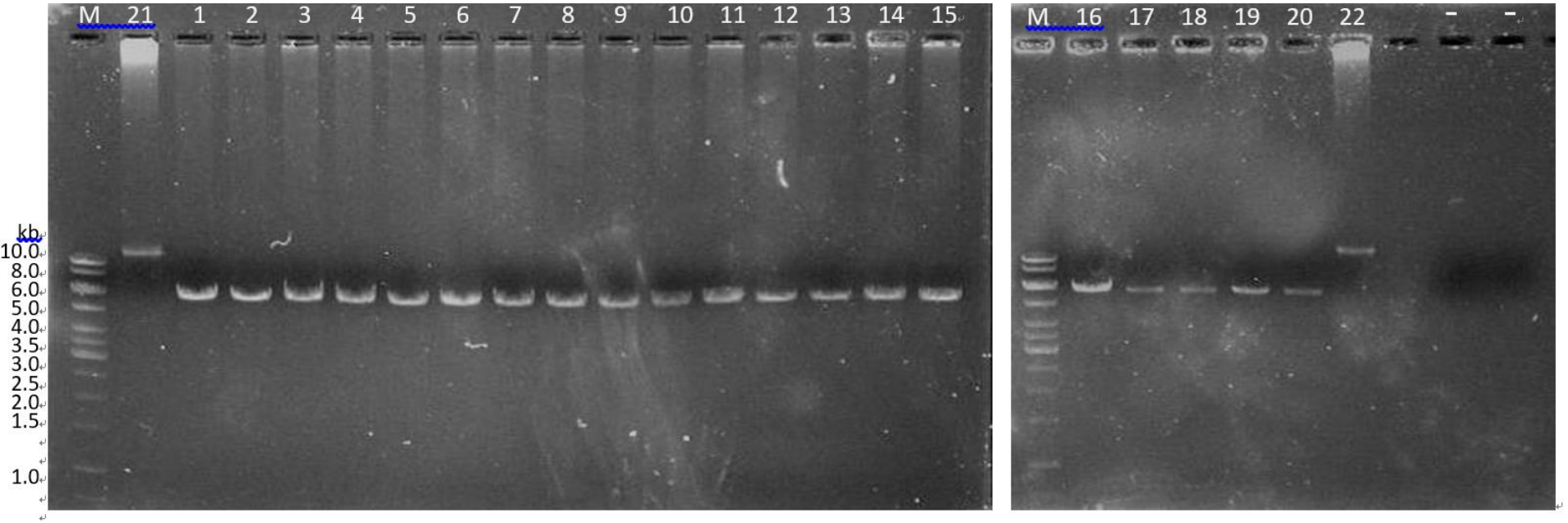
TWB			
CNV	**Mean glucose level**	**Lower.CL**	
Deletion	87.524	82.747	
Normal	91.977	91.819	
P-value: 0.067			

Validation of CNVs

Normal copy

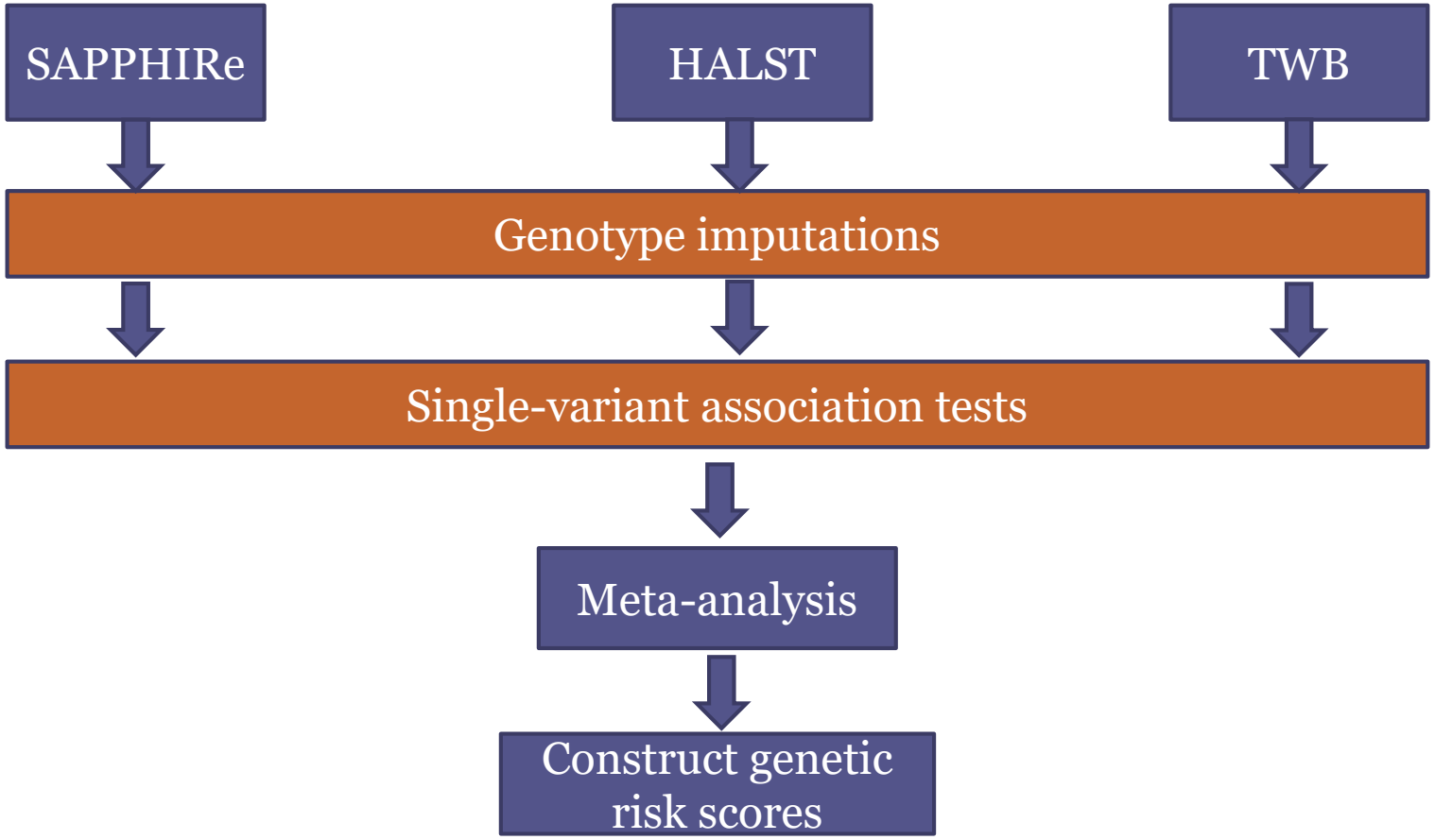
One copy deletion

Normal copy

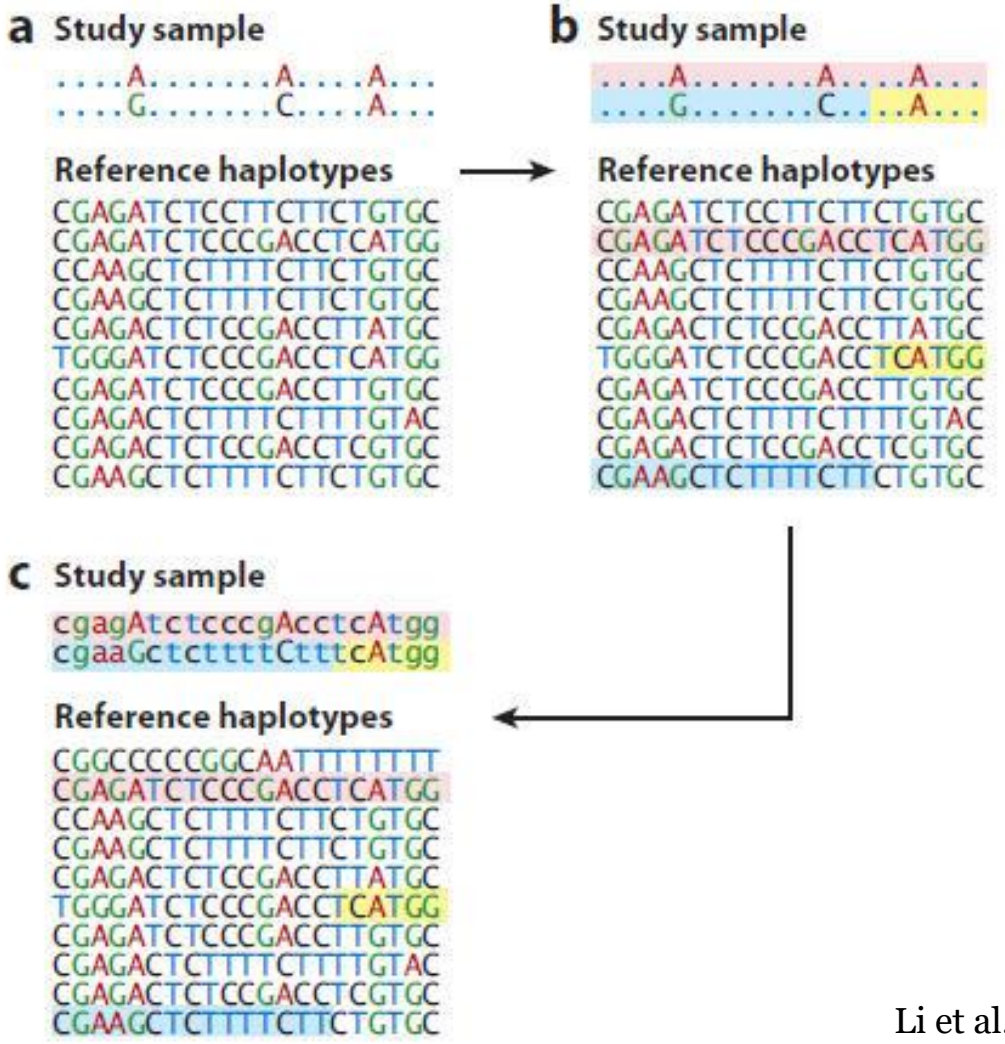


Expected length after deletion: 17.4kb – 10.96kb = 6.44kb

Single variant association analysis



Genotype imputation



Li et al. 2009. Annu Rev Genomics

LETTERS

nature genetics

Next-generation genotype imputation service and methods

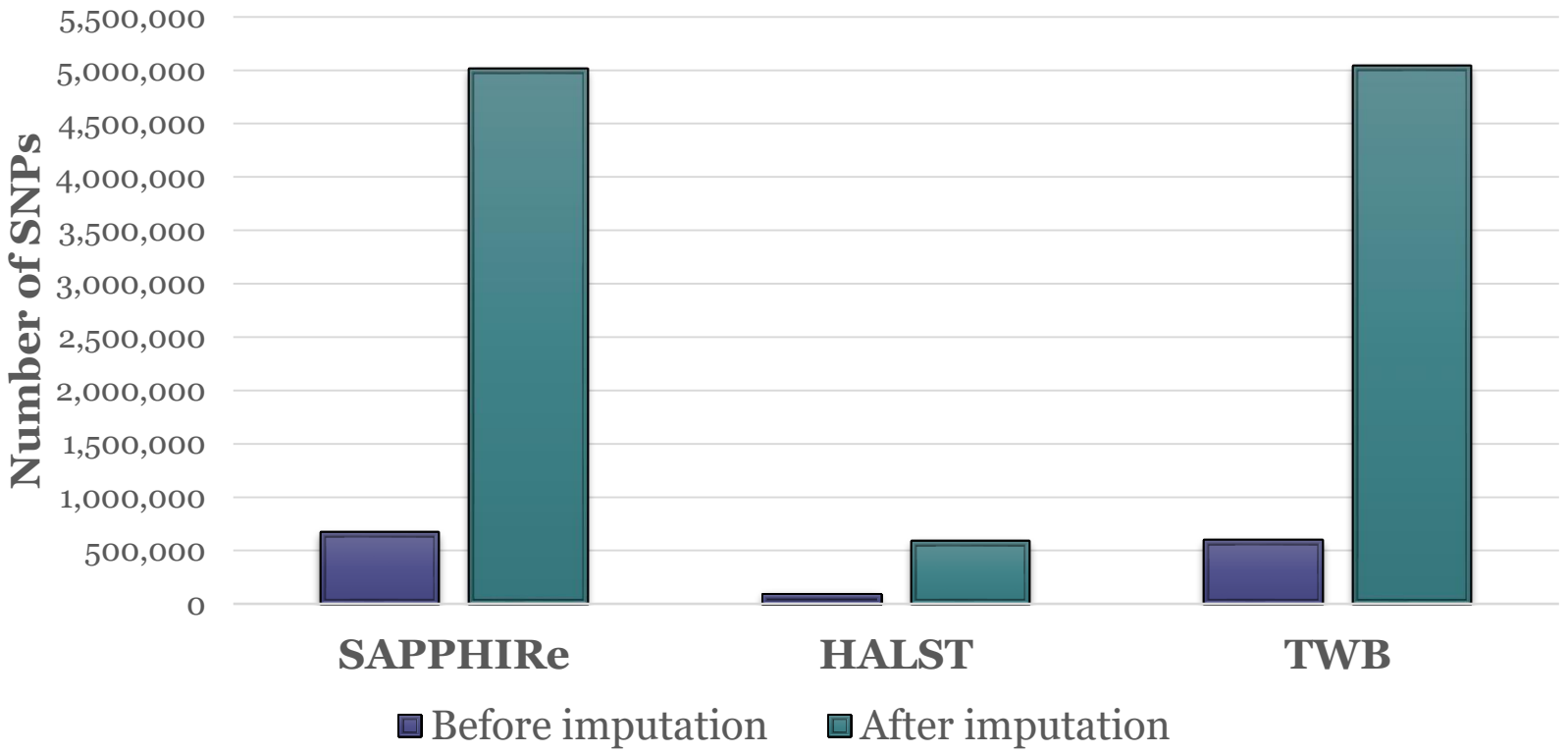
Sayantan Das[1,15], Lukas Forer[2,15], Sebastian Schönherr[2,15], Carlo Sidore[1,3,4], Adam E Locke[1], Alan Kwong[1], Scott I Vrieze[5], Emily Y Chew[6], Shawn Levy[7], Matt McGue[8], David Schlessinger[9], Dwight Stambolian[10], Po-Ru Loh[11,12], William G Iacono[8], Anand Swaroop[13], Laura J Scott[1], Francesco Cucca[3,4], Florian Kronenberg[2], Michael Boehnke[1], Gonçalo R Abecasis[1,16] & Christian Fuchsberger[1,2,14,16]

Genotype imputation is a key component of genetic association studies, where it increases power, facilitates meta-analysis, and aids interpretation of signals. Genotype imputation is computationally demanding and, with current tools, typically requires access to a high-performance computing cluster and to a reference panel of sequenced genomes. Here we describe improvements to imputation machinery that reduce computational requirements by more than an order of magnitude with no loss of accuracy in comparison to standard imputation tools. We also describe a new web-based service for imputation that facilitates access to new reference panels and

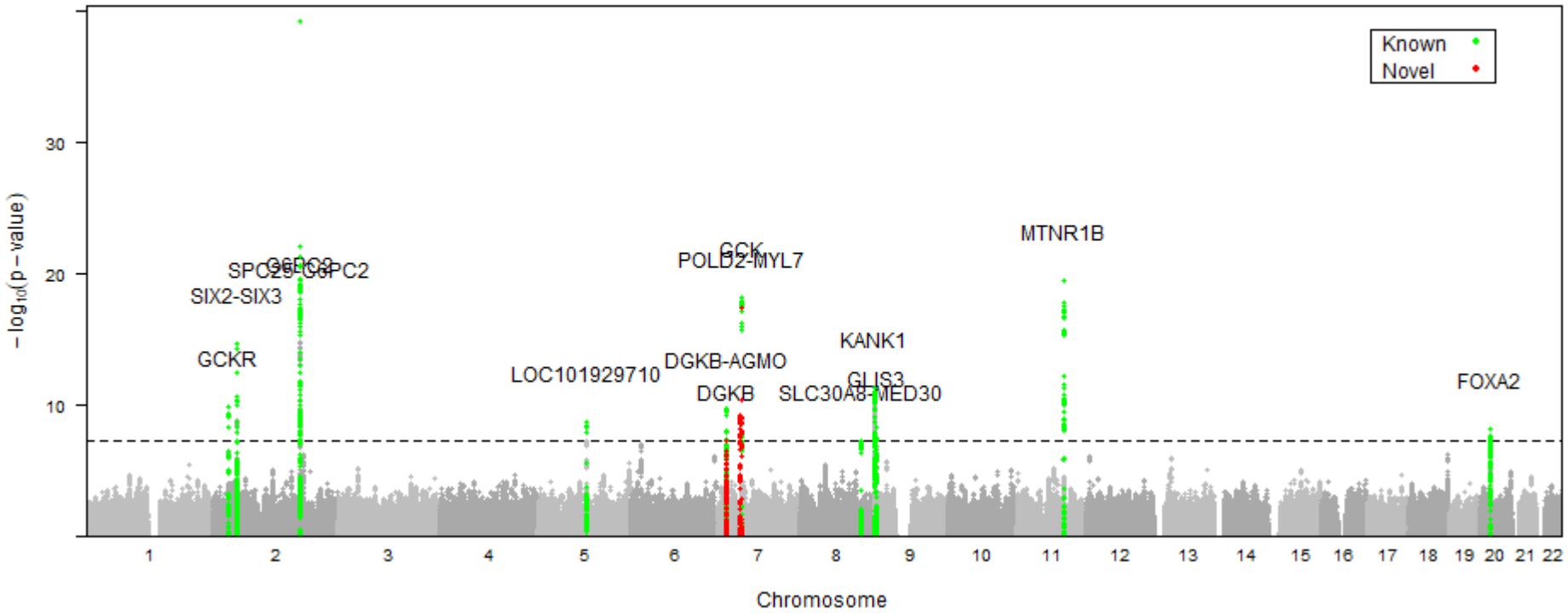
variants with frequencies of 0.1–0.5% or less and already include thousands of putative loss-of-function alleles. The HRC panel combines sequence data across >32,000 individuals from >20 medical sequencing studies and is cumbersome to access directly as a result of participant privacy protections in individual studies and the large volumes of data involved. Imputing 1,000 genome-wide association study (GWAS) samples using the HRC reference set requires ~2 years on a single CPU, or 1 week on a 100-core cluster, using minimac2 (ref. 8).

Here we present new algorithms for genotype imputation that increase computational efficiency with no loss of accuracy by leveraging local similarities between sequenced haplotypes. We also present

Genotype imputation

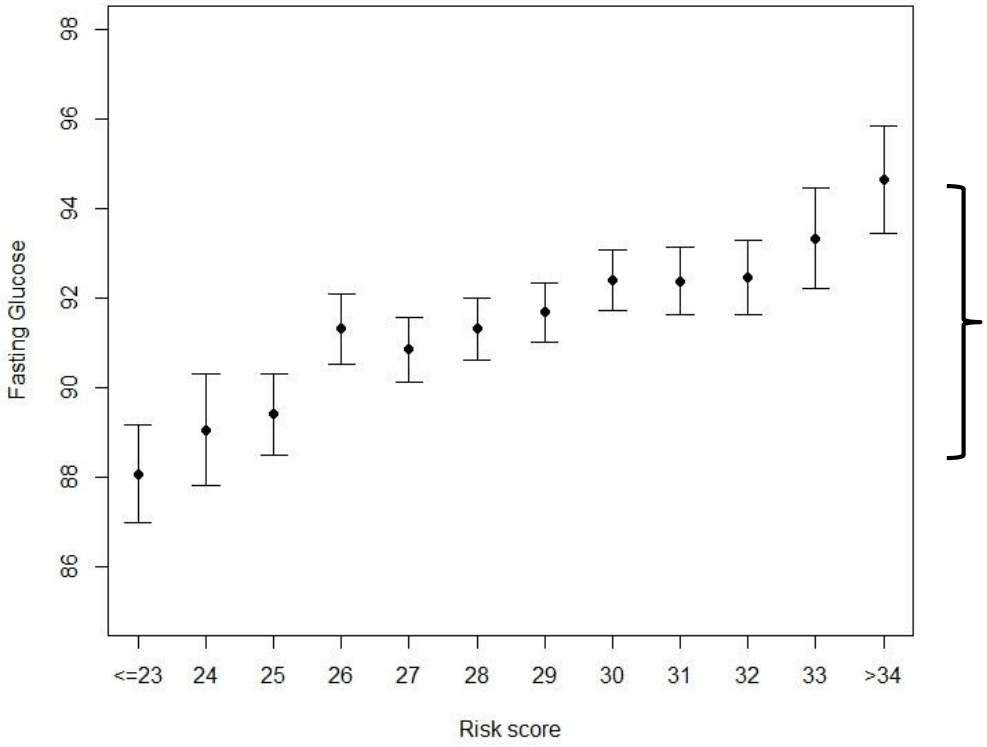


Manhattan plot for meta-analysis



Genetic risk scores

- 22 SNPs with high predictive power were used to construct genetic risk scores.



1.54-1.73 for future T2D

Gene-gene interaction analysis

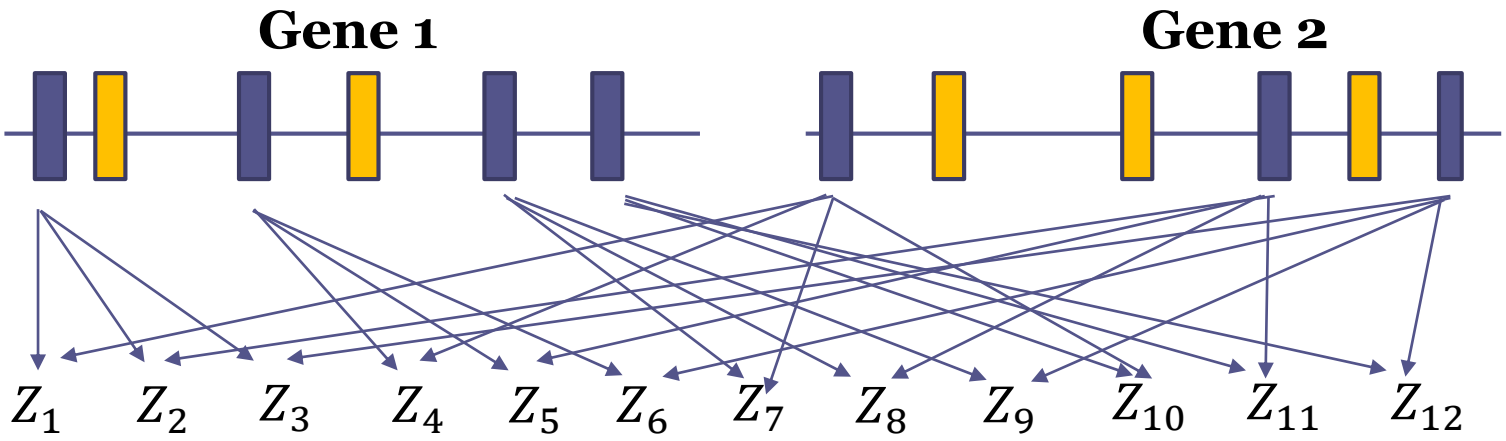
- Gene-gene interaction:

$$y = \alpha + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1 G_2$$

$$H_0: \beta_3 = 0$$

- Challenges: large number of tests
  - 20,000 genes have about 2×10^8 pairwise combinations
  - A modern computer needs more than 6 years to finish the analysis
  - A high multiple testing correction burden

Interaction test for Quantitative Trait (the IQT test)



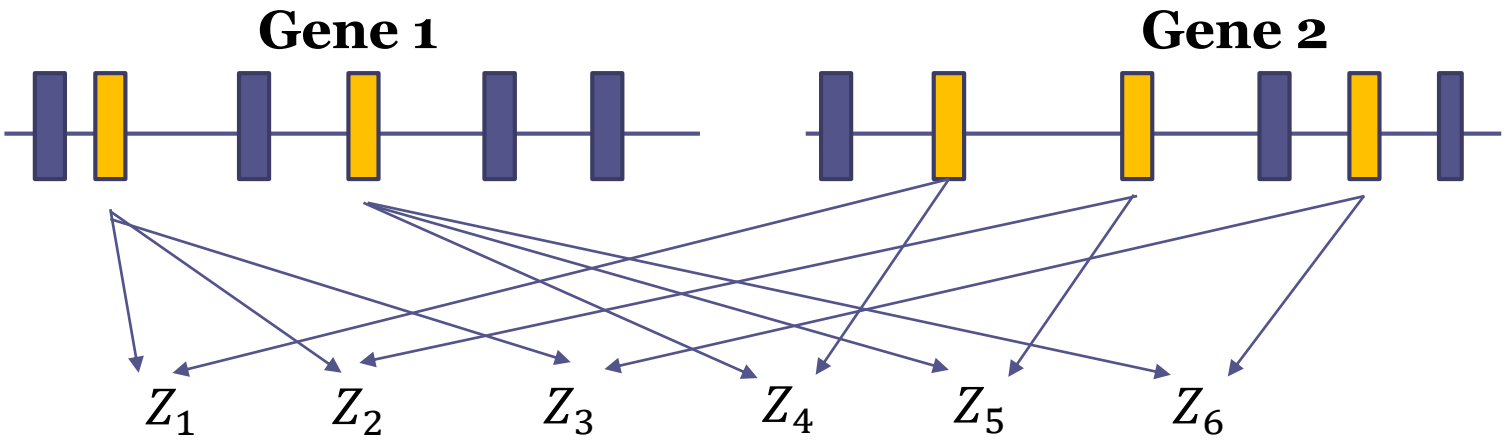
if p-value of $Z_i < 0.05$

$IQT_{common} = F(\boldsymbol{Z})$

Adaptive permutations

P-value of IQT_{common}

Interaction test for Quantitative Trait (the IQT test)



The top 5% of Z scores

$IQT_{rare} = F(\boldsymbol{Z})$

Adaptive permutations

P-value of IQT_{rare}

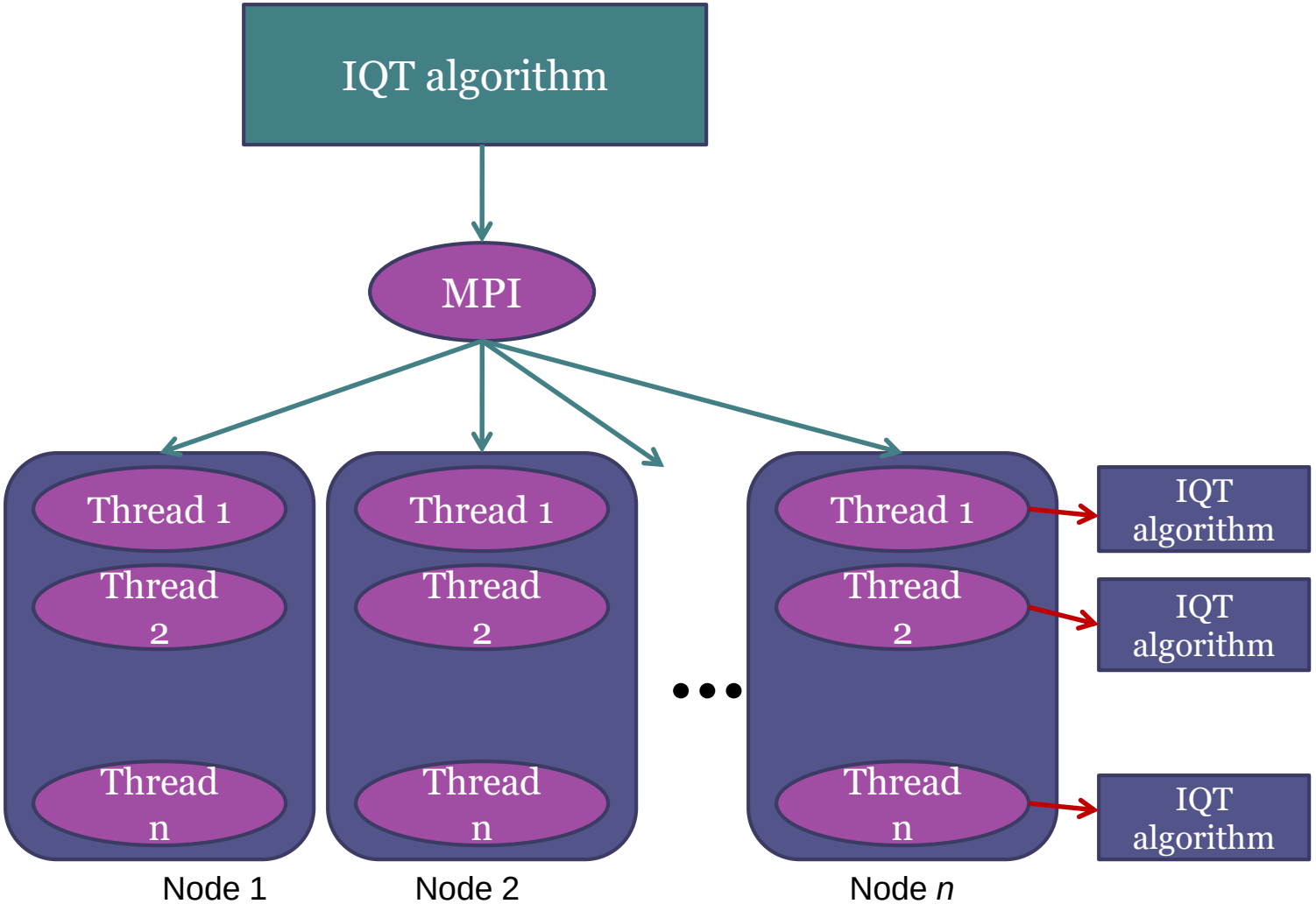
Performance improvement: A two-stage IQT test

- The distribution of the IQT statistic can be approximated by Wiens et al's method (2006):
  - For n independent standard normal random variables, the distribution of the sum of the k largest random variables, Y, can be simplified as:

$$\sqrt{n}\left(\frac{(Y/n)-\mu_\beta\left(\overline{F_n}\right)}{\sigma_\beta\left(\overline{F_n}\right)}\right) \sim N(0,1)$$

  - Permutations are performed only if the p-value based on the approximated distribution is $< m\alpha$.

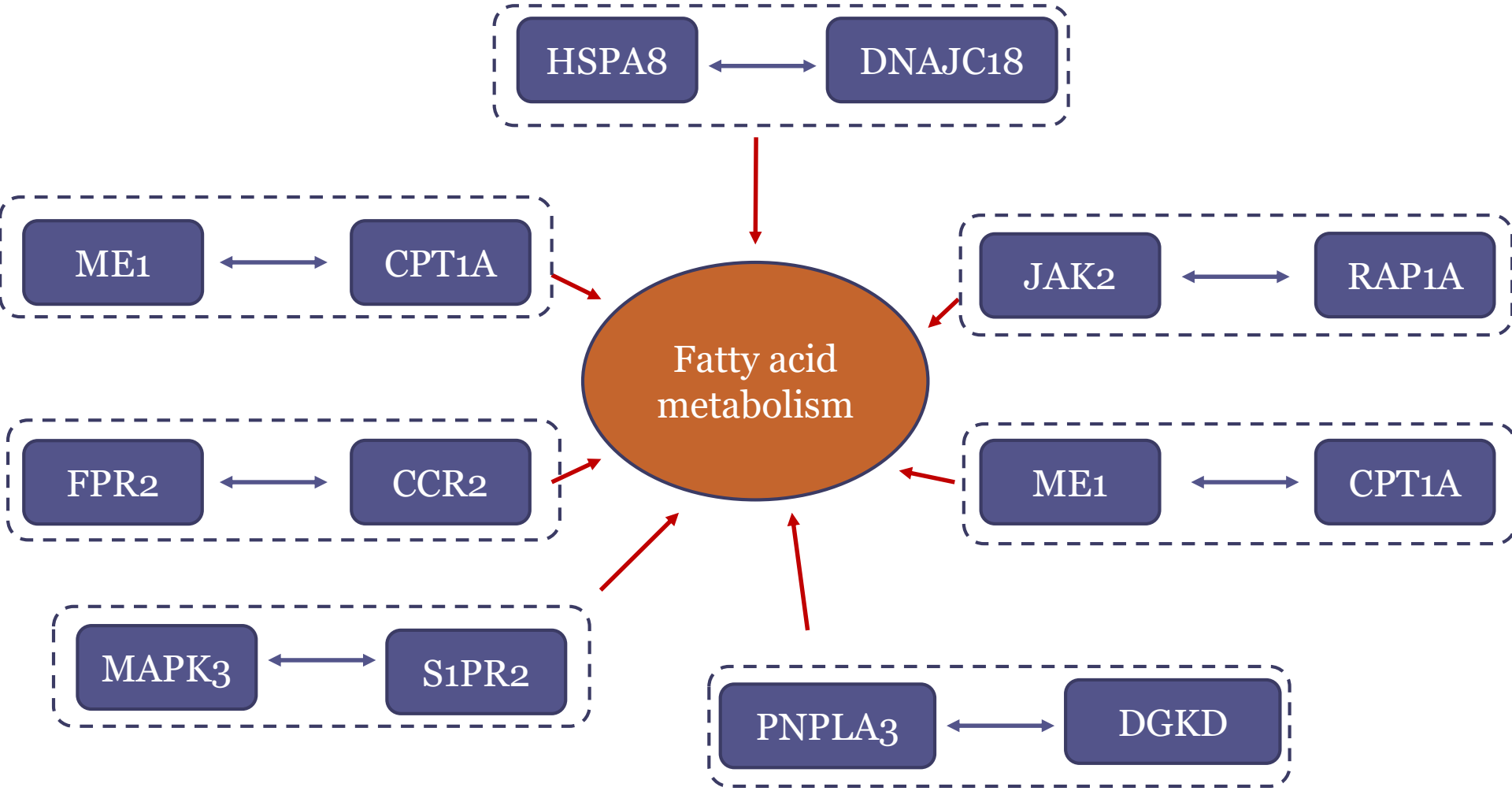
Performance improvement: MPI and threads



Performance improvement: biological knowledge-based method

- Studies have suggested that gene-gene interactions can be enriched in protein-protein interaction (PPI) networks.
- Protein-protein interaction database:
  - STRING
  - 542,895 PPI pairs with scores > 800 were downloaded.
- Gene-gene interaction tests were performed only in these PPI pairs.

Gene-gene interaction results



Conclusions

- The high-throughput genomics data have provided opportunities and challenges in data analysis.
- Our analyses based on CNVs, common SNPs, and gene-gene interactions have shown promising results for fasting glucose.
- An integrative approach for multi-omics data will become necessary.

Acknowledgements

- Collaborators:
  - Dr. Chao A. Hsiung
  - Dr. I-Shou Chang
  - Dr. Yen-Feng Chiu
- Research assistants:
  - Chen-Yu Kang
  - Kuo-Hong Lee
  - Ming-Che Dong