EVALUATION OF THE ASSOCIATION BETWEEN THE PATHOGENESIS OF TYPE 2 DIABETES AND GENOME-WIDE COPY NUMBER VARIATIONS USING THE LASSO METHOD

Wei Zhang[1,2], Yuan Ji[3,4†], Caihong Huang[5†], Jun Ying[6], Zhengqiang Ye[7], Guoyou Qin[8*], Naiqing Zhao[8*]

[1]Department of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China - [2]Fudan University Library, Fudan University, Shanghai 200032, China - [3]Northshore University HealthSystem, Evanston, IL 60201, USA - [4]University of Chicago Medicine and Biological Sciences, Chicago, IL 60637, USA - [5]Xuhui Central Hospital, Shanghai 200032, China - [6]Fudan University Library, Fudan University, Shanghai 200032, China - [7]Eye and ENT hospital of Fudan University, Shanghai 200032, China - [8]Department of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China

[†]*equal contribution*

ABSTRACT

***Objective**: Type 2 diabetes (T2D) is a complex disease caused by the combination of genetic factors and environmental factors. To date, although many loci, including genes and single nucleotide polymorphisms (SNPs), have been identified as risk variants of T2D, only approximately 10% of its heritability can be explained. In the current study, we proposed a data processing and analysis procedure to more accurately evaluate the association of the pathogenesis of T2D with copy number variations (CNVs).*

***Methods**: The data in our study came from the WTCCC (Wellcome Trust Case Control Consortium) genome-wide CNV database. Individual CNVs were identified by SW-ARRAY and CBS algorithms and genotyped with a global threshold method. Overlapped CNVs among all samples were split into smaller but more accurate CNV segments (CNVSegs) after the CNV call; then, LASSO-based logistic regression models with 10-fold cross-validations were performed 100 times to examine the association of CNVSegs with T2D. The AUC (area under the curve) in every model was summarized to preliminarily verify the classification ability of the models.*

***Results**: After quality control, 1,813 T2D cases and 2,777 controls were enrolled in the study. A total of 65,163 CNVs were identified, of which 25,512 were identified in the T2D group and 39,651 were identified in the healthy control group. A total of 22,279 CNVSegs were constructed after pre-processing the raw CNV data. By means of fitting 1,000 logistic regression models with the LASSO method, 26 CNVSegs were identified as T2D-associated CNVSegs according to pre-defined criteria (Frequency > 85% & Length > = 50 bp). Twenty-seven protein-coding genes were found to be overlapped with the CNVSegs, of which 11 were verified to be relevant to T2D, obesity or metabolic syndrome based on current published evidence. The average AUC of all models was 0.611 with the maximum being 0.683.*

***Conclusions**: Our study explored T2D-associated CNVSegs by LASSO-logistic regression models from the perspective of the whole genome for a more complete understanding of the genetic mechanisms of T2D. Further studies are necessary to verify the influence of the susceptibility loci on the pathogenesis or progression of T2D among different populations.*

***Keywords**: Type 2 diabetes, copy number variation, Genome-wide association, LASSO.*

DOI: 10.19193/0393-6384_2018_4_176

***Received** December 30, 2017; **Accepted** February 20, 2018*

Introduction

Diabetes is a chronic disease that occurs either when insulin cannot be sufficiently produced by the pancreas or effectively used by the body. In 2008, the age-standardized adult diabetes prevalence was 9.8% in men and 9.2% in women, and the number of diabetes patients was 347 million[(1)]. Type 2 diabetes (T2D), also called non-insulin-dependent diabetes, is a dominant subtype of diabetes that accounts for approximately 90% of people with diabetes around the world[(2)]. A combination of genetic factors and lifestyle was usually thought to be the main cause of the development of T2D[(3)].

In a population-based Framingham Offspring Study, age-adjusted odds ratio (95% CI) for offspring type 2 diabetes among individuals with maternal, paternal or bilineal diabetes were 3.4 (2.3-4.9), 3.5 (2.3-5.2) or 6.1 (2.9-13.0), respectively, relative to those without parental diabetes[4].

Generally, the first step in exploring the genetic basis was linkage and candidate gene studies, which were not very effective for T2D except for some rare familial forms[5]. SNP-based genome-wide association studies (GWAS) were a more powerful tool for detecting genetic susceptibility variants for complex diseases such as T2D because they do not make assumptions about disease pathogenesis[6]. Although many loci, including genes and single nucleotide polymorphisms (SNPs), were identified as risk variants of T2D, only approximately 10% of the heritability of T2D can be explaine[d(7-8)]. Therefore, the genetic architecture of T2D was considered to be rare variants in a common disease[9]. Additional studies are required to determine hidden or missing heritability and detect the causal variants within the identified loci for this disease. Apart from gene mutations and SNPs, copy number variations (CNVs) are also a major component of human genomic variation, which could influence gene expression, phenotypic variation and adaptation, and cause disease or confer risk to complex disease traits[10]. However, because CNVs are relatively new in examining the association of complex diseases and genetic markers, only a small number of T2D-associated CNVs have been revealed; for example, the LEPR gene locus, the region of the CAPN10 Indel19 marker, a CNV region (CNVR) located in chromosome 15 (chr15:45994758-45999227), as well as the finding of KCNIP1 as a modulator of insulin secretion[11-14].

In the current study, we propose a data processing and analysis procedure to identify copy number variants associated with T2D based on WTCCC (Wellcome Trust Case Control Consortium) samples[15]. Our method splits overlapped CNVs into smaller, but more precise CNVSegs after the CNV call. It then uses the LASSO-based logistic regression model to examine the association of CNVSegs with T2D in view of high-throughput traits of genome-wide CNV data. Our results were validated by the Database of Genomic Variants (DGV)[16] and 10-fold cross-validations 100 times. Genes in identified CNVSegs and their biological functions are also annotated in this paper.

Methods

Dataset

The data in our study were derived from the WTCCC copy number variation database which recorded information on the copy number variability of seven diseases including T2D. Hereditary information on 2,000 cases was included for each disease, as well as 3,000 shared controls from the UK 1958 birth cohort and the British Blood Center.

Identification of individual copy number variations and genotypes

The Affymetrix 500K SNP chip was employed as the CNV typing platform. The normalization scheme generated a single copy number intensity at each SNP from the multiple probe intensities for each allele. Copy number intensities at each SNP were converted to log2 ratios relative to the median intensity for that SNP on the same plate.

CNVs were identified by two segmental algorithms: SW-ARRAY[17] and Circular Binary Segmentation (CBS)[18]. CBS-generated segmentations (CNV calls) were retained only if they overlapped an SW-ARRAY CNV call. Samples with poor data quality were removed by the quality control procedure including SNP QC, segmentation QC and intensity QC. CNV calls were classified into deletion, copy-neutral LOH (loss-of-heterozygosity) and polysomy by lower and upper threshold values of 0.8 and 1.2, respectively, as suggested in the SW-ARRAY algorithm.

Data preprocessing and genetic biomarkers

In existing genome-wide CNV association studies, CNV regions (CNVRs) are often used as a genetic marker for association analysis. A CNVR defines a segment of DNA, with start and end loci, for which the copy number of the DNA segment is different from adjacent segments. CNVRs are called for single samples separately, and therefore different samples possess different break points. Most existing works merge break points across samples to form a unified CNVR. We took a different approach. Instead of merging, we split. That is, we construct shorter CNV segmentations (CNVSegs) by allowing all the break points across samples to form shorter segments. This split provides a higher resolution in our analysis. The difference between CNVR and CNVSegs is shown in Figure 1.

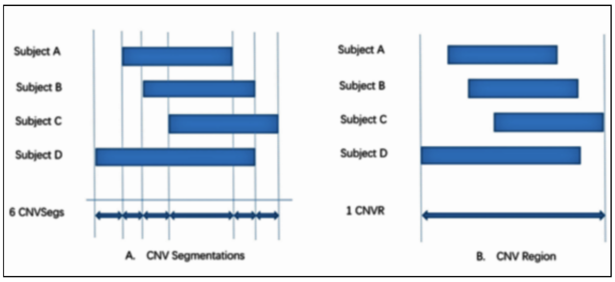


Figure 1: CNV segmentations vs. CNV regions. We assumed 4 subjects had overlapping CNVs in the genome. **A**. The overlapping boundaries of CNVs among 4 individuals were further subdivided to construct 6 CNV Segmentations; **B**. all CNVs among 4 individuals were merged into a CNV region.

Association analysis and validation

Conventional statistical analysis methods are not applicable to CNVSegs because of their high-throughput attribute. We have a large number of segmentations than sample size. We resort to a LASSO-based logistic regression model, in which age, sex and all CNVSegs were included as covariates. The LASSO method is a penalty regression in which both variable selection and model compression can be completed simultaneously. It provides shrinkage on parameter estimates and is designed to address high-dimensional and sparse data.

Due to the lack of external verification, we used 10-fold cross-validations repeated 100 times to ensure the robustness of the results (Figure 2).

One thousand fitted models were obtained and each model included CNVSegs that were more relevant to the diseased condition in the training set. CNVSegs with selection frequency greater than or equal to 85% and length greater than or equal to 50 bp were defined as T2D-associated CNVSegs. In addition to identifying disease-related high-frequency CNVSegs, we also summarized 1,000 AUC (area under the curve) values to preliminarily verify the classification ability of the models.

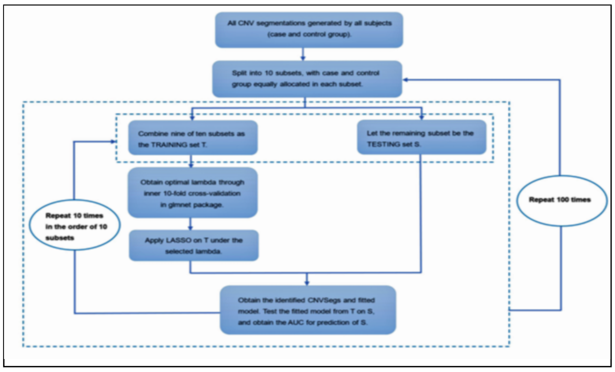

All CNV segmentations generated by all subjects (case and control group).
Split into 10 subsets, with case and control group equally allocated in each subset.
Combine nine of ten subsets as the TRAINING set T.
Let the remaining subset be the TESTING set S.
Obtain optimal lambda through inner 10-fold cross-validation in glmnet package.
Apply LASSO on T under the selected lambda.
Repeat 10 times in the order of 10 subsets
Repeat 100 times
Obtain the identified CNVSegs and fitted model. Test the fitted model from T on S, and obtain the AUC for prediction of S.

Figure 2: Flow diagram of LASSO-based logistic regression models with 10-fold cross-validations 100 times.

The following online databases were referenced to discover functional descriptions and genomic annotation of the genes overlapped with T2D-associated CNVSegs: Ensembl[(19)], Gencard[(20)], UCSC Genome Browser[(21)], Uniprot[(22)], and Genome Ontology[(23)].

Logistic regression with the LASSO method was implemented using the glmnet package[(24)] in R. To obtain the AUC of the fitted model in the validation set, the pROC package based on R was applied[(25)].

Results

After the quality control procedure, 1,813 T2D cases and 2,777 controls were included in the final analysis. The phenotype information is shown in Table 1.

	T2D group (n=1813)	Control group(n=2777)
Sex, No. (%)		
missing	0 (0)	163 (5.9)
male	1,048 (57.8)	1,289 (46.4)
female	765 (42.2)	1,325 (47.7)
Age, No. (%)		
missing	11 (0.6)	163 (5.9)
10-19	0 (0)	37 (1.3)
20-29	6 (0.3)	151 (5.4)
30-39	85 (4.7)	243 (8.8)
40-49	291 (16.0)	1,758 (63.3)
50-59	592 (32.7)	332 (12.0)
60-69	623 (34.4)	93 (3.3)
70-79	196 (10.8)	0 (0)
>/=80	9 (0.5)	0 (0)

Table 1: Characteristics of the subjects.

Characteristics of the identified CNVs

A total of 65,163 CNVs were identified, of which 25,512 were identified in the T2D group and 39,651 CNVs in the healthy control group. The distribution of average CNVs detected on different chromosomes was almost the same in the T2D and control groups (Figure 3). There were relatively more CNVs on chromosomes 14, 15 and 17, while there were fewer CNVs on chromosomes 13, 20 and 21. The average length of CNVs was also similar, which was 178 kb in the T2D group and 167 kb in the control group. The characteristics of genotyped CNVs are shown in Table 2.

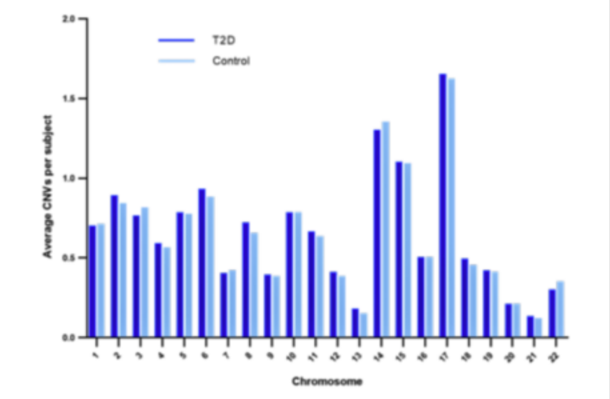


Figure 3: The distribution of average CNVs per subject identified on different chromosomes in different groups.

	Genotyped CNVs, n(%)			
	Total	Delete	Neutral LOH	Polysomy
T2D	25,512(100)	8,798(34.5)	10,487(41.1)	6,227(24.4)
Control	39,651(100)	13,960(35.2)	16,003(40.4)	9,688(24.4)

Table 2: The distribution of genotyped CNVs in different groups.

Characteristics of the generated CNVSegs

In the 4,590 subjects, 22,279 CNVSegs were obtained with our splitting scheme (Figure 1), of which 1,489 (6.68%) CNVSs were presented at frequencies greater than 1% and 479 (2.15%) were greater than 5% of the subjects. The distribution of CNVSegs on different chromosomes is shown in Figure 4.

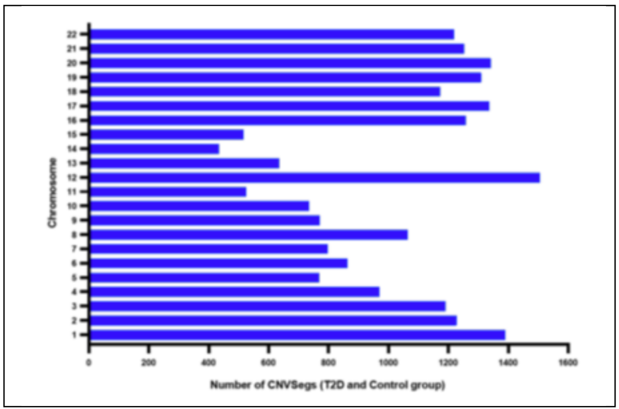
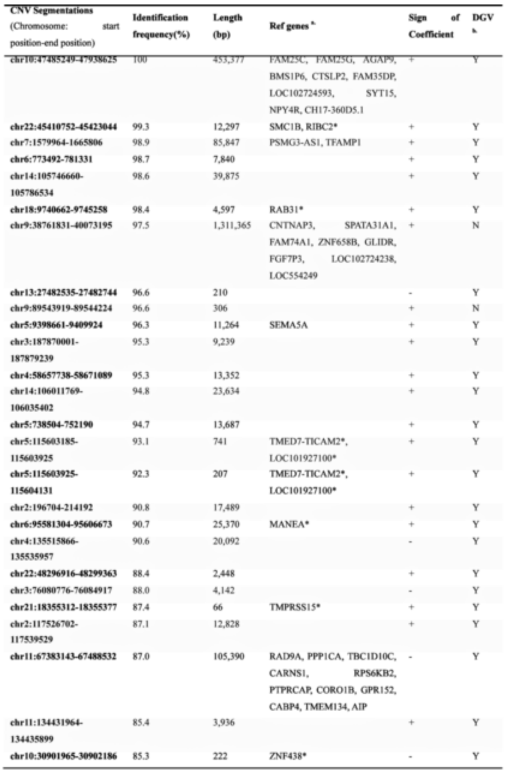


Fig. 4: The distribution of CNVSegs on different chromosomes in both groups.

LASSO-based logistic regression model with repeated cross-validations

A total of 1,894 CNVSegs were identified at least once by 1,000 LASSO-logistic models, of which 27 had selection frequencies over 85%. One CNVSeg less than 50 bp in length was excluded, and the remaining 26 CNVSegs were identified to be T2D-related CNVSegs, as defined in the "Methods" section. A total of 40 known genes were found to be overlapped with the CNVSegs, including 27 protein-coding genes, 7 RNA genes and 6 pseudo genes. Information about the T2D-related CNVSegs and overlapping genes is shown in Table 3. The functions and genomic annotations of the genes are shown in Table S1. Direct or indirect evidence showed that 11 genes (AGAP9, SYT15, NPY4R, Rab31, CNTNAP3, PPP1CA, MANEA, RPS6KB2, AIP, FAM25C and FAM25G) were relevant to T2D, obesity or metabolic syndrome.



CNV Segmentations (Chromosome: start position-end position)	Identification frequency(%)	Length (bp)	Ref genes [a]	Sign of Coefficient	DGV [b]
chr10:47485249-47938625	100	453,377	FAM25C, FAM25G, AGAP9, BMS1P6, CTSLP2, FAM35DP, LOC102724593, SYT15, NPY4R, CH17-360D5.1	+	Y
chr22:45410752-45423044	99.3	12,297	SMC1B, RIBC2*	+	Y
chr7:1579964-1665806	98.9	85,847	PSMG3-AS1, TFAMP1	+	Y
chr6:773492-781331	98.7	7,840		+	Y
chr14:105746660-105786534	98.6	39,875		+	Y
chr18:9740662-9745258	98.4	4,597	RAB31*	+	Y
chr9:38761831-40073195	97.5	1,311,365	CNTNAP3, SPATA31A1, FAM74A1, ZNF658B, GLIDR, FGF7P3, LOC102724238, LOC554249	+	N
chr13:27482535-27482744	96.6	210		-	Y
chr9:89543919-89544224	96.6	306		+	N
chr5:9398661-9409924	96.3	11,264	SEMA5A	+	Y
chr3:187870001-187879239	95.3	9,239		+	Y
chr4:58657738-58671089	95.3	13,352		+	Y
chr14:106011769-106035402	94.8	23,634		+	Y
chr5:738504-752190	94.7	13,687		+	Y
chr5:115603185-115603925	93.1	741	TMED7-TICAM2*, LOC101927100*	+	Y
chr5:115603925-115604131	92.3	207	TMED7-TICAM2*, LOC101927100*	+	Y
chr2:196704-214192	90.8	17,489		+	Y
chr6:95581304-95606673	90.7	25,370	MANEA*	+	Y
chr4:135515866-135535957	90.6	20,092		-	Y
chr22:48296916-48299363	88.4	2,448		+	Y
chr3:76080776-76084917	88.0	4,142		-	Y
chr21:18355312-18355377	87.4	66	TMPRSS15*	+	Y
chr2:117526702-117539529	87.1	12,828		+	Y
chr11:67383143-67488532	87.0	105,390	RAD9A, PPP1CA, TBC1D10C, CARNS1, RPS6KB2, PTPRCAP, CORO1B, GPR152, CABP4, TMEM134, AIP	-	Y
chr11:134431964-134435899	85.4	3,936		+	Y
chr10:30901965-30902186	85.3	222	ZNF438*	-	Y

Table 3: Information on the T2D-related CNVSegs and overlapping genes.

Comments: a. If the gene is marked by an asterisk, then CNV segmentation only covered a part of the gene. Otherwise, CNVSegs covered the whole gene.

b. Whether CNVSegs overlapped with the DGV database is shown.

Appendix A. Functional overview of known genes overlapped with T2D-related CNV segmentations

One thousand AUC values were obtained after 1,000 fitting processes. The average AUC was 0.611 with the maximum being 0.683.

Discussion

Diabetes is a chronic global epidemic caused by insufficient insulin secretion or the impaired biological function of insulin.

According to the latest report from the International Diabetes Federation, there were approximately 366 million diabetes patients worldwide, and this number will increase to 552 million in 2030. T2D in particular is a dominant subtype of diabetes, which significantly increases the risk of many serious health problems, such as heart attack, stroke, vision loss, and amputation. Therefore, it is meaningful and valuable to explore and clarify the etiology of T2D using existing genomics data.

Usually, a combination of some factors causes T2D, including genes, metabolic syndrome, extra weight, broken beta cells and others. Among them, heredity is significant in determining the risk of T2D because familial aggregation is one of the most important characteristics of T2D. The genetic model of T2D is thought to be a minor gene pattern (26); that is, there are no obvious critical genes among the co-acting genes. All of the co-acting genes are scattered at multiple loci and have a relatively lower frequency in the population, which then reaches the threshold of disease susceptibility or even the occurrence through interaction and the dose-effect relationship. Therefore, the majority of the heritability of T2D still remains missing, although recent large-scale genome-wide association studies have successfully identified some genetic loci associated with type 2 diabetes(27).

In this study, CNVSegs, as new genome-wide loci that can accurately measure the frequency of CNVs, were constructed to explore the genetic factors of T2D based on the large WTCCC data. Through the LASSO-based logistic regression model with repeated cross-validations, we identified 26 T2D-related CNVSegs overlapping with 40 known genes. Of the 27 protein-coding genes, 11 genes were found to be relevant to T2D, obesity or metabolic syndrome as presented by the following evidence.

AGAP9: An important homolog of this gene is ACAP4. A study in 2013(28) revealed that the phosphorylated ACAP4 mutant exhibited lipid-binding activity in vitro.

SYT15: In 2016, Sedova et al.(29) found that a limited genomic region on rat chromosome 16 significantly affected many features of metabolic syndrome from a novel congener rat model. The genes overlapping the chromosomal region just included SYT15, which was identified in the present study.

NPY4R: Aerts et al.(30) demonstrated that structural changes in the NPY4R gene played an important role in obesity through population-based CNV analysis and mutation screening in 2016.

Rab31: A 2012 study(31) showed that p75 (NTR), Rab5 and Rab31 regulate GLUT4 (glucose transporter 4) transport through a complex mechanism in adipocytes. GLUT4, in turn, regulates insulin-stimulated glucose uptake. Signaling pathways from p75 (NTR) to Rab5 or Rab31 may represent a unique therapeutic target for insulin tolerance and diabetes treatment.

CNTNAP3: Tews D et al.(32) certified that LRRC17, CNTNAP3, CD34, RGS7BP, and ADH1B were differentially expressed in brown adipocytes in a 2014 study. In another genome-wide analysis based on next-generation sequencing techniques(33), 6 genes including CNTNAP3, were significantly enriched in patients with obesity or diabetes.

PPP1CA: From the gene summary of NCBI, we learned that the protein encoded by this gene is one of the three catalytic subunits of protein phosphatase 1 (PP1). PP1 is a serine/threonine specific protein phosphatase known to be involved in the regulation of a variety of cellular processes, such as cell division, glycogen metabolism, muscle contractility, protein synthesis, and HIV-1 viral transcription.

MANEA: GO annotations related to this gene include alpha-mannosidase activity and glycoprotein endo-alpha-1,2-mannosidase activity.

TMPRSS15: In the gene summary of NCBI, this gene encodes an enzyme that converts the pancreatic proenzyme trypsinogen to trypsin, which activates other proenzymes including chymotrypsinogen and procarboxypeptidases.

RPS6KB2: Slattery et al.(34) suggested that RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 are involved in several pathways central to the carcinogenic process, including regulation of cell growth, insulin, and inflammation.

AIP: In a recent study(35), GH3-FTY mice (AIP gene knockout) showed insulin resistance relative to GH3 mice (AIP gene intact).

FAM25C and FAM25G: Both genes belong to the FAM25 family (family with sequence similarity 25). In 2016, WANG et al.(36) mentioned that six proteins, including FAM25, were involved in the redox reaction, energy metabolism of lipids and amino acid metabolism.

Although we found some T2D-associated genes using a new data processing and analysis procedure, there are also some shortcomings in our study. First, our data source was CNV data of a

European population requested and downloaded from the WTCCC. Therefore, although an internal validation method was applied, the identified genetic loci still need to be verified in an independent external population from our point of view. Second, in identification processes of individual CNV, the detection concordance of different algorithms applied to the same raw data is < 50% and reproducibility in replicate experiments is < 70% for most platforms(37). Therefore, the results of this study may partly be affected by the quality of current CNV detection platforms and corresponding algorithms. Finally, the classification ability using AUC as the indicator was limited in this study. This was due in part to the minor gene pattern in the genetics of T2D. Further studies are still needed to investigate the influence of other classification indicators or high-dimensional statistical analysis models on the results.

Conclusion

In this study, we explored copy number variations and their overlapping genes for T2D by presenting a data processing and analysis procedure. Our results may contribute to a more comprehensive understanding of the pathophysiology and genetic mechanism of T2D, but more in-depth studies are needed to further validate the impact of genetic variation found in this study on the pathogenesis or progression of type 2 diabetes in different populations.

References

1) Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet 2011; 378(9785): 31-40.

2) World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. In; 1999.

3) Murea M, Ma L, Freedman BI. Genetic and environmental factors associated with type 2 diabetes and diabetic vascular complications. Rev Diabet Stud 2012; 9(1): 6-22.

4) Meigs JB, Cupples LA, Wilson PW. Parental transmission of type 2 diabetes: the Framingham Offspring Study. Diabetes 2000; 49(12): 2201-2207.

5) Sanghera DK, Blackett PR. Type 2 Diabetes Genetics: Beyond GWAS. J Diabetes Metab 2012; 3(198).

6) Smushkin G, Vella A. Genetics of type 2 diabetes. Curr Opin Clin Nutr Metab Care 2010; 13(4): 471-477.

7) Herder C, Roden M. Genetics of type 2 diabetes: pathophysiologic and clinical relevance. Eur J Clin Invest 2011; 41(6): 679-692.

8) Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012; 44(9): 981-990.

9) Grarup N, Sandholt CH, Hansen T, Pedersen O. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. Diabetologia 2014; 57(8): 1528-1541.

10) Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature 2006; 444(7118): 444-454.

11) Jeon JP, Shim SM, Nam HY, Ryu GM, Hong EJ, Kim HL, et al. Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. BMC Genomics 2010; 11: 426.

12) Plengvidhya N, Chanprasert K, Tangjittipokin W, Thongnoppakhun W, Yenchitsomanus PT. Identification of copy number variation of CAPN10 in Thais with type 2 diabetes by multiplex PCR and denaturing high performance liquid chromatography (DHPLC). Gene 2012; 506(2): 383-386.

13) Bae JS, Cheong HS, Kim JH, Park BL, Kim JH, Park TJ, et al. The genetic effect of copy number variations on the risk of type 2 diabetes in a Korean population. PLoS One 2011; 6(4): e19091.

14) Lee HS, Moon S, Yun JH, Lee M, Hwang MY, Kim YJ, et al. Genome-wide copy number variation study reveals KCNIP1 as a modulator of insulin secretion. Genomics 2014; 104(2): 113-120.

15) WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447(7145): 661-678.

16) MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 2014; 42(Database issue): D986-992.

17) Price TS, Regan R, Mott R, Hedman A, Honey B, Daniels RJ, et al. SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. Nucleic Acids Res 2005; 33(11): 3455-3464.

18) Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004; 5(4): 557-572.

19) Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res 2016; 44(D1): D710-716.

20) Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 1997; 13(4): 163.

21) Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser

database: 2017 update. Nucleic Acids Res 2017; 45(D1): D626-D634.

22) Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. Methods Mol Biol 2017; 1558: 41-55.

23) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25(1): 25-29.

24) Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33(1): 1-22.

25) Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77.

26) Bao MX. Predisposing genes of the onset of type 2 diabetes mellitus. Chin J Clin Rehabil 2006; 10(16): 140-143.

27) Kato N. Insights into the genetic basis of type 2 diabetes. J Diabetes Investig 2013 ;4(3): 233-244.

28) Zhao X, Wang D, Liu X, Liu L, Song Z, Zhu T, et al. Phosphorylation of the Bin, Amphiphysin, and RSV161/167 (BAR) domain of ACAP4 regulates membrane tubulation. Proc Natl Acad Sci U S A 2013; 110(27): 11023-11028.

29) Sedova L, Pravenec M, Krenova D, Kazdova L, Zidek V, Krupkova M, et al. Isolation of a Genomic Region Affecting Most Components of Metabolic Syndrome in a Chromosome-16 Congenic Rat Model. PLoS One 2016; 11(3): e0152708.

30) Aerts E, Beckers S, Zegers D, Van Hoorenbeeck K, Massa G, Verrijken A, et al. CNV analysis and mutation screening indicate an important role for the NPY4R gene in human obesity. Obesity (Silver Spring) 2016; 24(4): 970-976.

31) Baeza-Raja B, Li P, Le Moan N, Sachs BD, Schachtrup C, Davalos D, et al. p75 neurotrophin receptor regulates glucose homeostasis and insulin sensitivity. Proc Natl Acad Sci U S A 2012; 109(15): 5838-5843.

32) Tews D, Schwar V, Scheithauer M, Weber T, Fromme T, Klingenspor M, et al. Comparative gene array analysis of progenitor cells from human paired deep neck and subcutaneous adipose tissue. Mol Cell Endocrinol 2014; 395(1-2): 41-50.

33) Crespo-Facorro B, Prieto C, Sainz J. Schizophrenia gene expression profile reverted to normal levels by antipsychotics. Int J Neuropsychopharmacol 2014; 18(4).

34) Slattery ML, Lundgreen A, Herrick JS, Wolff RK. Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. Mutat Res 2011; 706(1-2): 13-20.

35) Fukuda T, Tanaka T, Hamaguchi Y, Kawanami T, Nomiyama T, Yanase T. Augmented Growth Hormone Secretion and Stat3 Phosphorylation in an Aryl Hydrocarbon Receptor Interacting Protein (AIP)-Disrupted Somatotroph Cell Line. PLoS One 2016; 11(10): e0164131.

36) Wang Y, Kou Y, Wang X, Cederbaum A, Wang R. Multifactorial comparative proteomic study of cytochrome P450 2E1 function in chronic alcohol administration. PLoS One 2014; 9(3): e92504.

37) Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 2011; 29(6): 512-520.

Acknowledgement
YJ's research is partly supported by NIH 2R01 CA132897.

Availability of data and materials
The data that support the findings of this study are available from WTCCC but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of WTCCC.

Corresponding author
NAIQING ZKAO
Department of Biostatistics, School of Public Health, Fudan University
Shanghai 200032
E mail: nqzhao@fudan.edu.cn
(China)
GUOYOU QIN
Department of Biostatistics, School of Public Health, Fudan University
Shanghai 200032
E mail: gyqin@fudan.edu.cn
(China)