Exome-chip association analysis reveals an Asian-specific missense variant in *PAX4* associated with type 2 diabetes in Chinese individuals

Chloe Y. Y. Cheung¹ · Clara S. Tang² · Aimin Xu^{3,4,5} · Chi-Ho Lee¹ · Ka-Wing Au¹ · Lin Xu⁶ · Carol H. Y. Fong¹ · Kelvin H. M. Kwok¹ · Wing-Sun Chow¹ · Yu-Cho Woo¹ · Michele M. A. Yuen¹ · JoJo S. H. Hai¹ · Ya-Li Jin⁷ · Bernard M. Y. Cheung¹ · Kathryn C. B. Tan¹ · Stacey S. Cherny⁸ · Feng Zhu⁷ · Tong Zhu⁷ · G. Neil Thomas⁹ · Kar-Keung Cheng⁹ · Chao-Qiang Jiang⁷ · Tai-Hing Lam^{6,7} · Hung-Fat Tse^{1,10} · Pak-Chung Sham^{8,11,12} · Karen S. L. Lam^{1,3,4}

Received: 4 May 2016 / Accepted: 21 September 2016 / Published online: 15 October 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract

Aims/hypothesis Genome-wide association studies (GWASs) have identified many common type 2 diabetes-associated variants, mostly at the intronic or intergenic regions. Recent advancements of exome-array genotyping platforms have opened up a novel means for detecting the associations of low-frequency or rare coding variants with type 2 diabetes. We conducted an exomechip association analysis to identify

additional type 2 diabetes susceptibility variants in the Chinese population.

Methods An exome-chip association study was conducted by genotyping 5640 Chinese individuals from Hong Kong, using a custom designed exome array, the Asian Exomechip. Single variant association analysis was conducted on 77,468 single nucleotide polymorphisms (SNPs). Fifteen SNPs were subsequently genotyped for replication analysis in an independent

Chloe Y. Y. Cheung and Clara S. Tang contributed equally to this work and are joint first authors.

Karen S. L. Lam, Pak-Chung Sham, Hung-Fat Tse and Tai-Hing Lam contributed equally to the supervision of this work and are joint corresponding authors.

Electronic supplementary material The online version of this article (doi:10.1007/s00125-016-4132-z) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

✉ Tai-Hing Lam
hmrhth@hku.hk

✉ Hung-Fat Tse
hftse@hkucc.hku.hk

✉ Pak-Chung Sham
pcsham@hku.hk

✉ Karen S. L. Lam
ksllam@hku.hk

¹ Department of Medicine, University of Hong Kong, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, People's Republic of China

² Department of Surgery, University of Hong Kong, Hong Kong, People's Republic of China

³ State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, Hong Kong, People's Republic of China

⁴ Research Centre of Heart, Brain, Hormone and Healthy Ageing, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, People's Republic of China

⁵ Department of Pharmacology & Pharmacy, University of Hong Kong, Hong Kong, People's Republic of China

⁶ School of Public Health, Room 505, Faculty of Medicine Building, William M.W. Mong Block, University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong, People's Republic of China

⁷ Molecular Epidemiological Research Centre, Guangzhou Number 12 Hospital, Guangzhou, Guangdong, People's Republic of China

⁸ Department of Psychiatry, University of Hong Kong, Hong Kong, People's Republic of China

Chinese cohort comprising 12,362 individuals from Guangzhou. A combined analysis involving 7189 cases and 10,813 controls was performed.

Results In the discovery stage, an Asian-specific coding variant rs2233580 (p.Arg192His) in *PAX4*, and two variants at the known loci, *CDKN2B-AS1* and *KCNQ1*, were significantly associated with type 2 diabetes with exome-wide significance ($p_{\text{discovery}} < 6.45 \times 10^{-7}$). The risk allele (T) of *PAX4* rs2233580 was associated with a younger age at diabetes diagnosis. This variant was replicated in an independent cohort and demonstrated a stronger association that reached genome-wide significance ($p_{\text{meta-analysis}} [p_{\text{meta}}] = 3.74 \times 10^{-15}$) in the combined analysis.

Conclusions/interpretation We identified the association of a *PAX4* Asian-specific missense variant rs2233580 with type 2 diabetes in an exome-chip association analysis, supporting the involvement of *PAX4* in the pathogenesis of type 2 diabetes. Our findings suggest *PAX4* is a possible effector gene of the 7q32 locus, previously identified from GWAS in Asians.

Keywords Asian-specific · Exome-chip association analysis · *PAX4* · Type 2 diabetes

Abbreviations

CSN1S1	Casein α S1
Chr	Chromosome
FGF21	Fibroblast growth factor 21
FGFR1	Fibroblast growth factor receptor 1
FPG	Fasting plasma glucose
GBCS	Guangzhou Biobank Cohort Study
GWAS	Genome-wide association studies
HKU-TRS	University of Hong Kong Theme-based Research Scheme
HWE	Hardy–Weinberg equilibrium
LD	Linkage disequilibrium
MAF	Minor allele frequency
PAX	Paired box
PC	Principal component
QC	Quality control

SGLT1	Solute carrier family 5 member 1
SNP	Single nucleotide polymorphism
TTBK2	Tau tubulin kinase 2

Introduction

Type 2 diabetes is a common disease resulting from the complex interactions between multiple genetic and environmental factors. Insights into the genetic basis of type 2 diabetes will facilitate the discovery of novel treatment targets. Since 2007, the success in genome-wide association studies (GWAS) has led to the identification of a large number of independent loci for type 2 diabetes. However, the disease-susceptibility single nucleotide polymorphisms (SNPs) identified from these GWAS are common variants which tend to confer relatively small effect sizes, altogether accounting for only 10–15% of the type 2 diabetes heritability [1]. The functional consequences of these susceptibility variants, which are mostly present in intronic or intergenic regions, remained difficult to interpret. In the past few years, the role of low-frequency (minor allele frequency [MAF] = 1–5%) and rare (MAF < 1%) coding variants with various complex traits [2–7] is being increasingly studied. It has been suggested that most current rare variants were introduced by mutational events during the recent explosive growth of the human population [8, 9]. These rare variants are believed to confer a greater effect than the common variants because of the limited time for purifying selection to act [9, 10]. The majority of efforts to reveal type 2 diabetes susceptibility variants have been made in populations of European ancestry. Using advanced technologies, such as the exome chip and whole-genome/exome sequencing, researchers have detected associations of additional novel coding variants, both common and rare, for type 2 diabetes [4, 5] and several quantitative glycaemic traits, such as fasting glucose and insulin levels, in European populations [3, 6, 7]. As samples of European ancestry represent only a subset of human genetic variations [11], the risk variants in other populations are likely to be insufficiently characterised. A genome-wide trans-ancestry meta-analysis reported several type 2 diabetes-susceptibility variants which showed significant differences in effect sizes and associations in different populations [12]. For instance, the effect size of *TCF7L2* rs7903146 was higher in Europeans than in East Asians, whilst the association signal of *PEPD* rs3786897 was specific to the populations of East Asians and the association signal of *KLF14* rs13233731 was only significant in the European samples [12]. Such observations highlight the importance of conducting association analyses in non-European populations to detect novel loci affecting the risk of type 2 diabetes.

The advancement in array-based genotyping technology, such as exome arrays, has provided a more cost-effective approach than whole-genome or exome sequencing for assessing

⁹ Institute of Applied Health Research, University of Birmingham, Birmingham, UK

¹⁰ Hong Kong-Guangdong Joint Laboratory on Stem Cell and Regenerative Medicine, University of Hong Kong, Hong Kong, People's Republic of China

¹¹ Centre for Genomic Sciences, Centre for Genomic Sciences, University of Hong Kong, 6/F, HKJC Building for Interdisciplinary Research, 5 Sassoon Road, Pokfulam, Hong Kong, People's Republic of China

¹² State Key Laboratory in Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, People's Republic of China

the association of rare and low-frequency coding variants that may be population specific. In a joint collaborative study, our group has recently reported several novel or Asian-specific coding variants associated with blood lipids [2] using a tailored Illumina HumanExome BeadChip (Asian Exomechip [13]). In the present study, we aimed to detect novel loci for type 2 diabetes in the Chinese population using this Asian Exomechip. We first conducted an exome-chip association analysis based on 5640 participants from the University of Hong Kong Theme-based Research Scheme (HKU-TRS) cohort, and genotyped 15 SNPs for replication in an independent Southern Han Chinese cohort from Guangzhou ($n = 12,362$).

Methods

Participants

Discovery cohort The discovery stage involved a total of 5640 Southern Han Chinese participants (3652 cases and 1988 controls) from the HKU-TRS cohort who participated in a previous exome-chip association study for blood lipid traits [2]. The study participants were recruited from the Hong Kong West Diabetes Registry (HKWDR) [14], the Hong Kong Cardiovascular Risk Factor Prevalence Study (CRISPS) [15] and the Hong Kong Chinese coronary artery disease (CAD) cohort. All participants were recruited from these clinic-based or community-based studies conducted at the Queen Mary Hospital, Hong Kong, People's Republic of China. Details of the corresponding cohorts have been previously reported [2]. Type 2 diabetes cases were defined as meeting at least one of the following criteria: fasting plasma glucose (FPG) ≥ 7 mmol/l, 2 h glucose during OGTT ≥ 11.1 mmol/l, taking glucose-lowering agents or physician-diagnosed diabetes. All controls had no documented history of diabetes and were not receiving treatment for diabetes. Written informed consent was obtained from each participant and the study protocol was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster.

Replication cohort The replication stage involved a total of 12,362 Southern Han Chinese participants (3537 cases and 8825 controls) of the Guangzhou Biobank Cohort Study (GBCS). The clinical characteristics and glycaemic status (outlined in the electronic supplementary material [ESM], Table 1) were based on cross-sectional data obtained at the time of blood sample collection. Details of the GBCS have been described previously [16], however, in brief, the GBCS is a collaborative project between Guangzhou No. 12 Hospital, the University of Hong Kong and the University of Birmingham (Birmingham, UK). The GBCS was established to examine the effect of genetic and environmental influences on health problems and the development of chronic diseases.

Baseline recruitment was conducted from 2003 to 2008 ($n = 30,519$; age, ≥ 50 years) in Guangzhou [17]. Participants were invited to have a second examination from August 2008 to December 2012. The present study included participants who attended the second examination and from whom sufficient information was obtained to determine type 2 diabetes status. Type 2 diabetes cases were defined as meeting at least one of the following criteria: FPG ≥ 7 mmol/l, 2 h glucose during OGTT ≥ 11.1 mmol/l, HbA_{1c} $\geq 6.5\%$ (≥ 47.5 mmol/mol), taking glucose-lowering agents or self-reported physician-diagnosed diabetes. All controls were not receiving treatment for diabetes, had no documented history of diabetes and had FPG < 6.1 mmol/l and 2 h glucose during OGTT < 7.8 mmol/l. Written informed consent was obtained from each participant and the study protocol was approved by the Guangzhou Medical Ethics Committee of the Chinese Medical Association.

Genotyping and data quality control

Discovery stage All participants were genotyped using the Asian Exomechip, which is a specially designed exome array with an add-on content of 58,317 variants in addition to the standard content of the Infinium HumanExome BeadChip (HumanExome-12v1_A; Illumina, San Diego, CA, USA). A detailed description of the Asian Exomechip design has been presented elsewhere [2, 13]; briefly, the standard content of the exome array includes 242,901 markers, including: (1) $>200,000$ protein-altering variants identified from approximately 12,000 sequenced genomes and exomes of primarily European ancestry; (2) $>20,000$ non-exonic variants contributed by multiple consortia, such as the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project; (3) variants designed for ancestry differentiation, sample tracking and for establishing segments of identity by descent (http://genome.sph.umich.edu/wiki/Exome_Chip_Design; accessed 1 May 2016). The European-based design of the Infinium HumanExome BeadChip has led to an under-representation of non-European genomes and thereby limited the coverage of low-frequency variants among non-European populations. To allow comprehensive genotyping across the full allele frequency spectrum in Asians, a custom panel of $\sim 30,000$ missense or nonsense variants identified from three independent Asian sequencing datasets of ~ 1000 Chinese samples were integrated into the Asian Exomechip. Additionally, a custom set of common variants selected for GWAS follow-up or fine mapping studies was also included. Genotype calling was conducted by GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina). We first conducted manual inspection of genotype clusters for $>55,000$ variants that had either a GenTrain score < 0.8 , high missingness ($> 1\%$), or a poor genotype clustering, as determined by exome-chip genotyping of > 9000 individuals by collaborators [13, 18]. A total of 4550 variants with poor genotype clustering were removed. Individual-level quality control

(QC) was conducted with regard to sex mismatch, duplication, biological relatedness and possible sample contamination. A principal component (PC) analysis was conducted to examine the existence of non-Chinese samples using a panel of >20,000 independent common SNPs (MAF > 0.05), with outliers excluded from the analysis. For SNP-level QC, we excluded 217,455 SNPs with MAF < 0.1%, of which 179,107 SNPs were monomorphic, 154 SNPs deviated from Hardy–Weinberg equilibrium (HWE) with $p < 1 \times 10^{-5}$ in controls, 3854 SNPs had >2% missingness and 8086 SNPs were originally designed for the purpose of QC (including the fingerprint SNPs for sample tracking, ancestry informative markers [AIMs] for distinguishing Europeans from native and African-Americans, and grid SNPs for the identification of identity by descent segments). After QC measures, a total of 5640 participants and 77,468 variants were included in the association analysis.

Replication stage In the replication stage, we genotyped all SNPs that achieved $p_{\text{discovery}} < 5 \times 10^{-4}$ and with potential functional relevance, except *CDKN2B-AS1/DMRTA1* rs10965250 and *KCNQ1* rs2237896, which have been previously reported to be of genome-wide significance ($p < 5 \times 10^{-8}$) in GWAS [1] and also reached exome-wide significance ($p_{\text{discovery}} < 6.45 \times 10^{-7}$) in the current study. By SNPs with potential functional relevance, we refer to SNPs at or near genes that showed protein–protein interactions, shared a pathway with known type 2 diabetes susceptibility genes or were implicated in the pathogenesis of diabetes. These included *PAX4* rs2233580, *CDKAL1* rs10440833, *FGFR1* rs2288696, *ANKRD55/MAP3K1* rs456867, *IGF2BP2* rs11711477, *TTBK2* rs56017612 and *DUSP26/UNC5D* rs4739563, *HCG27/HLA-C* rs3869115, *SCN1B* rs67701503, *DAP* rs267939, *CSN1S1* rs10030475, *ZNF283/ZNF404* rs138993781, *STAB1* rs740903, *CARNS1* rs868167 and *PDPN* chromosome (Chr)1:13937002. All 15 selected SNPs were then genotyped using the MassARRAY Sequenom platform (San Diego, CA, USA) at the Beijing Genomics Institute (BGI), Beijing, People's Republic of China. Four SNPs that either showed a low genotyping call rate (<90%; *PDPN* Chr1:13937002 and *IGF2BP2* rs11711477) or deviated from HWE in controls ($p_{\text{HWE}} < 0.003 = 0.05/15$; *CDKAL1* rs10440833 and *SCN1B* rs67701503) in the replication study were excluded from further analysis. Thus, for the final analysis, a total of 11 SNPs were included. *PAX4* rs2233580 did not deviate from HWE in controls and was therefore retained in the analysis, even though it significantly deviated from HWE in the case group ($p_{\text{HWE}} < 0.003$). *PAX4* has been implicated in the pathogenesis of type 2 diabetes [19], and it is recognised that a true association can lead to deviation from HWE in cases [20]. The average genotyping call rate of these 11 SNPs was 98.2%.

Variants annotation and in silico functional analysis

Function of variants and protein changes for non-synonymous SNPs were annotated by KGGSeq (version 1.0; <http://grass.cgs.hku.hk/limx/kggseq/>; accessed 15 April 2016) [21] according to the RefGene annotation. The pathogenic potential of the non-synonymous variants was assessed through various deleteriousness and conservation prediction tools implemented in KGGSeq, including SIFT [22] and PolyPhen [23].

Asian-specific variants

Variants were classified as 'Asian-specific' if they were monomorphic in both the European and African populations but polymorphic (MAF > 0) in the Asian population, according to the 1000 Genomes Project [11].

Data analysis

All statistical analyses in the discovery and replication stages were conducted using PLINK, version 1.9 (<http://pngu.mgh.harvard.edu/~purcell/plink/>) [24]. In the discovery stage, multiple logistic regression analysis with adjustment for age, sex and the first two PCs was employed to examine for the associations with type 2 diabetes under the additive genetic model. To assess the adiposity independent association of the top SNPs (with $p_{\text{discovery}} < 5 \times 10^{-4}$) with type 2 diabetes, we further included BMI in the multiple logistic regression model. Exome-wide significance was defined as $p < 6.45 \times 10^{-7}$ ($=0.05/77,468$). To address the between-SNP linkage disequilibrium (LD), the p value-informed LD-based clumping approach, with the '–clump' command implemented in PLINK, was conducted. The index SNP had the most significant p value from each clumped association region. Each index SNP formed clumps with other variants which were in LD with the index SNP ($r^2 \geq 0.2$) and were within ± 500 kb from the index SNP. The association between *PAX4* rs2233580 and age of diabetes diagnosis was examined by univariate linear regression analysis. In the replication stage, age and sex were included as covariates in the multiple logistic regression model to assess for associations with type 2 diabetes. A Bonferroni corrected one-tailed p value $< 4.54 \times 10^{-3}$ ($=0.05/11$) was used as the threshold for successful replication. Meta-analysis of the association results of the discovery and replication stages were conducted using METAL (<http://csg.sph.umich.edu/abecasis/Metal/>; accessed 1 May 2016) [25]. The inverse variance fixed-effect method was employed to pool the summary statistics of the two stages and heterogeneity of effect was assessed using Cochran's Q test and I^2 index.

Table 1 Clinical characteristics of study participants in the discovery stage

Variable	Controls	Cases	<i>p</i> value
Number	1988	3652	-
Male (%)	56.4	60.9	<0.001
Age (years)	58.7 ± 12.1	64.8 ± 11.8	<0.001
Fasting glucose (mmol/l)	5.1 ± 0.6	7.6 ± 2.5	<0.001
BMI (kg/m ²)	24.2 ± 3.7	25.9 ± 4.0	<0.001
Waist circumference (cm)			
Male	86.3 ± 8.4	91.5 ± 10.2	<0.001
Female	79.1 ± 9.1	86.3 ± 10.9	<0.001
Coronary artery disease (%)	37.2	43.8	<0.001
Hypertension (%)	40.3	85.5	<0.001
Use of anti-hypertensive drug (%)	31.5	83.0	<0.001
Use of lipid-lowering drug (%)	37.5	65.6	<0.001
Ever smoker (%)	34.4	35.9	0.251

Data are mean ± SD

Results

A total of 5640 Chinese (Hong Kong) participants (see Table 1 for details) were genotyped using a custom Asian Exomechip. Single-variant association analysis was performed to assess the associations with type 2 diabetes for 77,468 polymorphic variants (Fig. 1). Of these, 48% altered protein composition and 21% were Asian-specific variants with a MAF between 0.1% and 5%.

In the discovery stage, single-variant association analysis was conducted in 3652 cases and 1988 controls, adjusted for age, sex and the first two PCs. We detected 34 index SNPs within 32 loci significantly associated with type 2 diabetes at

$p_{\text{discovery}} < 5 \times 10^{-4}$ (ESM Table 2), of which three variants reached exome-wide significance ($p_{\text{discovery}} < 6.45 \times 10^{-7}$) (Table 2). These included the known associations at *CDKN2B-AS1/DMRTA1* rs10965250 ($p_{\text{discovery}} = 5.93 \times 10^{-8}$, OR [95% CI] 0.80 [0.74, 0.87]) and *KCNQ1* rs2237896 ($p_{\text{discovery}} = 1.82 \times 10^{-7}$; OR [95% CI] 0.80 [0.73, 0.87]) reported in previous GWAS, as well as an Asian-specific variant, rs2233580 (p.Arg192His) of *PAX4* ($p_{\text{discovery}} = 1.75 \times 10^{-7}$; OR [95% CI] 1.39 [1.23, 1.56]). As *PAX4* is a known gene for MODY [26], we further examined its association with age of diabetes diagnosis. The risk allele (T) of rs2233580 was found to be significantly associated with younger age of diabetes diagnosis ($p = 6.01 \times 10^{-4}$ [ESM Table 3]; β [95% CI] −1.45 [−2.28, −0.62] [data not shown]; mean age of diagnosis ± SD [years] TT 52 ± 13, CT 53 ± 13, CC 54 ± 13 [ESM Table 3]). In addition, we identified several loci not previously reported to be associated with type 2 diabetes: *FGFR1* rs2288696 ($p_{\text{discovery}} = 2.29 \times 10^{-5}$; OR [95% CI] 0.73 [0.63, 0.85]), *TTC22* rs56017612 ($p_{\text{discovery}} = 7.40 \times 10^{-5}$; OR [95% CI] 0.72 [0.61, 0.84]) and *DUSP26/UNC5D* rs4739563 ($p_{\text{discovery}} = 7.48 \times 10^{-5}$; OR [95% CI] 0.80 [0.72, 0.90]) (ESM Table 2). The association of all SNPs with type 2 diabetes remained statistically significant after further adjustment for BMI (ESM Table 2).

In the replication stage, 11 of the 15 selected SNPs passed QC and were analysed in 3537 cases and 8825 controls. Replication and combined association results of these SNPs are shown in Table 3. Of these SNPs, eight showed consistent direction of effects. Only the association of the *PAX4* missense variant, rs2233580, with type 2 diabetes was successfully replicated (one-tailed $p_{\text{replication}} = 1.22 \times 10^{-9}$; OR [95% CI] 1.28 [1.18, 1.39]; remained significant after Bonferroni correction). Meta-analysis of the association results gave a genome-wide

Fig. 1 Manhattan plot of discovery stage results. The y-axis represents the $-\log_{10}(p \text{ value})$, and the x-axis represents the genomic position. The dots represent the 77,468 SNPs analysed, relative to their position on each

chromosome (alternating black and grey). The black horizontal dashed line indicates exome-wide significance (6.45×10^{-7}). Diamonds show the exome-wide significant SNPs

Table 2 Association results of SNPs reaching exome-wide significance ($p < 6.45 \times 10^{-7}$) in the discovery stage

Nearest gene(s)	SNP	Position	Annotation	A1	A2	MAF		OR (95% CI)	$p_{\text{discovery}}^{\text{a}}$	$p_{\text{discovery}}^{\text{b}}$
						Cases	Controls			
Asian-specific variant										
<i>PAX4</i>	rs2233580	7:127253550	p.Arg192His	T	C	0.145	0.113	1.39 (1.23, 1.56)	1.75×10^{-7}	7.62×10^{-6}
Established type 2 diabetes susceptibility variants										
<i>CDKN2B-AS1/DMRTA1</i>	rs10965250	9:22133284	Intergenic	A	G	0.384	0.430	0.80 (0.74, 0.87)	5.93×10^{-8}	8.80×10^{-10}
<i>KCNQ1</i>	rs2237896	11:2858440	Intronic	A	G	0.311	0.359	0.80 (0.73, 0.87)	1.82×10^{-7}	1.53×10^{-8}

A1, minor allele; A2, major allele ORs are relative to the minor allele

^a Adjusted for age, sex, PC1 and PC2

^b Adjusted for age, sex, BMI, PC1 and PC2

significant association, with no evidence of heterogeneity in effect size ($p_{\text{meta}} = 3.74 \times 10^{-15}$, OR [95% CI] 1.31 [1.23, 1.40] [Table 3]; $I^2 = 10$, $p_{\text{heterogeneity}} = 0.292$ [data not shown]). The associations of *FGFR1* rs2288696, *TTBK2* rs56017612 and *DUSP26/UNC5D* rs4739563 were not significant in the replication cohort. However, the direction of effects for both *FGFR1* rs2288696 and *TTBK2* rs56017612 were consistent with those from the discovery stage. A modest association was observed at a missense variant of *CSN1S1* (rs10030475 [p.Pro137Thr]; one-tailed $p_{\text{replication}} = 7.5 \times 10^{-3}$, OR [95% CI] 0.93 [0.87, 0.99]). However, this association did not pass Bonferroni correction for multiple testing in the replication stage.

Discussion

The present study reports the first exome-chip association analysis on type 2 diabetes in a Chinese population. By genotyping 5640 Chinese participants using a custom Asian Exomechip, which interrogated 77,468 polymorphic SNPs, we identified the association of an Asian-specific coding variant in *PAX4* and replicated the associations of some known type 2 diabetes-susceptibility loci. We also detected a few possible candidates which showed potential functional relevance in the pathogenesis of type 2 diabetes, such as *TTBK2*, *FGFR1* and *CSN1S1*.

Table 3 Replication and combined association results

Gene(s)	SNP	A1	Discovery Hong Kong (3652 cases vs 1988 controls)		Replication Guangzhou (3537 cases vs 8825 controls)		Combined Hong Kong + Guangzhou (7189 cases vs 10,813 controls)		
			OR (95% CI)	$p_{\text{discovery}}^a$	OR (95% CI)	One-tailed $p_{\text{replication}}^a$	Dir	OR (95% CI)	p_{meta}^a
<i>PAX4</i>	rs2233580	T	1.39 (1.23, 1.56)	1.75×10^{-7}	1.28 (1.18, 1.39)	1.22×10^{-9b}	++	1.31 (1.23, 1.40)	3.74×10^{-15}
<i>FGFR1</i>	rs2288696	A	0.73 (0.63, 0.85)	2.29×10^{-5}	0.98 (0.88, 1.09)	0.350	--	0.88 (0.81, 0.96)	4.57×10^{-3}
<i>ANKRD55/MAP3K1</i>	rs456867	T	0.84 (0.77, 0.91)	4.78×10^{-5}	0.99 (0.93, 1.05)	0.363	--	0.94 (0.89, 0.98)	8.57×10^{-3}
<i>TTBK2</i>	rs56017612	C	0.72 (0.61, 0.84)	7.40×10^{-5}	0.90 (0.80, 1.02)	0.046	--	0.83 (0.75, 0.92)	2.11×10^{-4}
<i>DUSP26/UNC5D</i>	rs4739563	T	0.80 (0.72, 0.90)	7.48×10^{-5}	1.00 (0.93, 1.08)	0.518	+-	0.93 (0.87, 0.99)	0.020
<i>HCG27/HLA-C</i>	rs3869115	C	0.81 (0.73, 0.90)	1.04×10^{-4}	0.99 (0.92, 1.07)	0.418	--	0.93 (0.87, 0.99)	0.016
<i>DAP</i>	rs267939	A	0.79 (0.69, 0.89)	1.85×10^{-4}	1.01 (0.92, 1.10)	0.545	+-	0.92 (0.86, 0.99)	0.035
<i>CSN1S1</i>	rs10030475	T	0.85 (0.78, 0.92)	1.86×10^{-4}	0.93 (0.87, 0.99)	7.50×10^{-3}	--	0.90 (0.85, 0.94)	3.28×10^{-5}
<i>ZNF283/ZNF404</i>	rs138993781	G	0.27 (0.13, 0.55)	2.93×10^{-4}	0.69 (0.34, 1.39)	0.148	--	0.43 (0.26, 0.71)	1.03×10^{-3}
<i>STAB1</i>	rs740903	T	1.19 (1.08, 1.31)	3.02×10^{-4}	1.00 (0.94, 1.07)	0.450	++	1.06 (1.01, 1.12)	0.030
<i>CARNS1</i>	rs868167	A	0.67 (0.54, 0.83)	3.33×10^{-4}	1.05 (0.90, 1.23)	0.730	+-	0.90 (0.79, 1.02)	0.109

ORs are relative to the minor allele.

^a Adjusted for age and sex

^b Retained significance following Bonferroni correction for multiple testing in the replication analysis

For effects in the same direction as in the discovery stage analysis, one-tailed p values were calculated as $p/2$; for effects in the opposite direction as in the discovery stage analysis, one-tailed p values were calculated as $1 - p/2$

A1, minor allele; Dir, direction of effect

The identification of the Asian-specific and probably damaging variant of *PAX4* is the major finding of this study. *PAX4* encodes a member of the paired box (PAX) family of paired-homeodomain factors. *PAX4* functions as a transcription repressor and plays a crucial role in pancreatic beta cell function and development [27]. It also plays a role in beta cell proliferation and survival [28, 29]. Heterozygous *Pax4*-knockout mice harbour less mature pancreatic beta and delta cells, but have numerous abnormally clustered alpha cells, indicating the essential role of *PAX4* in the differentiation of beta and delta cell lineages [30]. *PAX4* has been shown to repress the transcriptional activity of insulin [19] and glucagon [31] promoters. *PAX4* is located at 7q32, a region reported to be associated with type 2 diabetes in previous GWAS of Asians [32, 33]. An intergenic variant rs6467136 located near *GRIP* and *GCC1-PAX4* was reported to be associated with type 2 diabetes in a meta-analysis of eight GWAS in East Asians [32], whilst rs10229583, located downstream of *PAX4*, was identified as a risk variant in a GWAS for type 2 diabetes in a Chinese population [33]. Such observations, together with our findings, suggest that the effect of *PAX4* may be more evident in East Asians than in other populations. The association of type 2 diabetes with both of these SNPs appear to be independent of the missense variant rs2233580 (p.Arg192His), which was identified in the current study. According to the 1000 Genomes project [11], rs2233580 shows very low LD with both rs6467136 ($r^2=0.03$) and rs10229583 ($r^2=0.02$). The association of rs6467136 was not significant ($p_{\text{discovery}}=0.284$). Data for rs10229583 were not available for analysis in the present study. Our findings provide evidence that *PAX4* is a possible effector gene at 7q32, a GWAS locus for type 2 diabetes. Our exome chip achieves 50% coverage of the coding variants within this gene region. Nonetheless, we were unable to eliminate the possibility that the association of rs2233580 with type 2 diabetes that was identified in the current study resulted from tagging of other causative coding variants which were not covered by our exome chip. However, its functional significance, as demonstrated by in silico [22, 23] and in vitro [34, 35] analyses, suggests that this SNP is likely to be the causative variant. While in silico analysis of the two previously reported intergenic variants was unable to define their functional relevance (RegulomeDB score = 5 for rs10229583 and = 6 for rs6467136), rs2233580 was predicted to be damaging by multiple prediction tools (SIFT score = 0; PolyPhen2 HDIV score = 1; Polyphen2 HVAR score = 0.99) [22, 23]. An in vitro study showed that the transcriptional repressor activities of *PAX4* p.Arg192His on human insulin and glucagon promoters were reduced when compared with wild-type *PAX4* [34]. The Arg192 residue is highly conserved across different species, including human, mouse, rat and chimpanzee, and this residue has been shown to make direct contact with the major groove of DNA-binding sequences [35]. An amino acid change in the homeodomain of *PAX4* may cause a defect in its transcriptional activity. It has been proposed that this variant may affect diabetes risk through its effect on beta cell proliferation in the

adult pancreas, or beta cell differentiation and maturation during development, leading to beta cell mass reduction [34]. While rs2233580 has a frequency of ~10% among Asian populations, according to the 1000 Genomes project this variant was found to be monomorphic in European and African individuals [11], suggesting interrogation in less-studied non-European populations would facilitate the identification of novel population-specific associations. Our finding of an Asian-specific variant also has implications for the construction of polygenic genetic scores to predict type 2 diabetes in Asian populations.

Our observation of the significant association of *PAX4* rs2233580 with type 2 diabetes was in agreement with findings from a large-scale whole-genome/exome sequencing study conducted by the GoT2D and T2D-GENES consortia [36], which was recently published during the review process of our manuscript. rs2233580 was reported to be associated with type 2 diabetes exclusively in 2165 East Asian individuals at genome-wide significance ($p=9.3 \times 10^{-9}$), and this association was further replicated in three independent East Asian cohorts [36]. Mutations in *PAX4* have been found to cause the rare monogenic form, MODY, in Thai individuals [26]. On the other hand, common variants of a number of established MODY genes have been found to be associated with type 2 diabetes, including *GCK*, *HNF1 α* (also known as *HNF1A*), *HNF4 α* (also known as *HNF4A*), *HNF1 β* (also known as *HNF1B*) and *PDX1* [37–39]. Findings from the current study and those reported by the GoT2D and T2D-GENES consortia suggest that *PAX4* also harbours common variants that confer susceptibility to type 2 diabetes. Interestingly, in a previous GWAS of East Asians, the risk allele of a common variant, rs10229583, located downstream of *PAX4*, was reported to be associated with higher risk of type 2 diabetes and a younger age of diagnosis [33]. Among the 3652 cases in the current study, individuals who carried the risk allele (T) of the *PAX4* missense variant rs2233580 were also significantly younger at the time of diagnosis. In contrast, the GoT2D and T2D-GENES consortia reported no significant association between rs2233580 and age of diagnosis in a total of 1619 cases from three independent cohorts of East Asian ancestry (Hong Kong Chinese, Korean and Singapore Chinese) [36]. This contradictory observation could be attributed to the much larger sample size of the current study, which provided sufficient power to detect the association (ESM Table 3). Furthermore, study heterogeneity caused by different ascertainment criteria for type 2 diabetes cases in the studies may have also contributed to the discordant observations. A meta-analysis of our data with those of the three independent cohorts has provided evidence to support the association of *PAX4* rs2233580 with younger age at diagnosis ($p_{\text{meta}}=0.007$; z score = -2.717 ; $I^2=58.5$, $p_{\text{heterogeneity}}=0.065$; ESM Table 3).

Although unable to reach genome/exome-wide significance, the potential functions of *TBK2*, *FGFR1* and *CSN1S1* have made them possible candidates for type 2

diabetes. Tau tubulin kinase 2 (TTBK2) is a serine/threonine kinase known to phosphorylate tau and tubulin [40]. TTBK2 is involved in regulation of the sodium-dependent glucose transporter, solute carrier family 5 member 1 (SGLT1) [41], which is responsible for the absorption of glucose and galactose in the intestine and is involved in the reabsorption of glucose in the kidney [42]. Depletion of TTBK2 has been shown to decrease SGLT1 stability in the cell membrane and lead to loss of glucose transport capacity in *Xenopus* oocytes [41]. Mice with attenuated fibroblast growth factor receptor 1 (FGFR1) signalling exhibited a reduced number of beta cells, impaired expression of glucose transporter 2, enhanced proinsulin content in beta cells and developed diabetes with age [43]. FGFR1 is the primary receptor of fibroblast growth factor 21 (FGF21) and hence regulates FGF21 responsiveness. FGF21 has shown beneficial metabolic effects in animals and humans [44] and our team previously demonstrated that high FGF21 levels could predict type 2 diabetes development [15]. The paradoxical increase in FGF21 levels in patients with type 2 diabetes suggest that FGF21 resistance may play a role in the pathogenesis of type 2 diabetes [44]. Our finding that a variant of *FGFR1* is associated with type 2 diabetes is supportive of such a possibility. Casein α S1 (CSN1S1) is a member of the casein family that has been shown to possess proinflammatory properties, such as the upregulation of IL-1 β [45]. Data from animal studies and clinical trials is suggestive of a causative role for IL-1 β in the loss of beta cell mass in type 2 diabetes [46]. Overall, given their potential functional relevance in the pathogenesis of type 2 diabetes, more detailed investigation of these genes, such as deep sequencing analysis, is warranted.

A potential limitation of the present study was the underrepresentation of rare functional variants specific to the Chinese populations in the exome chip. With an attempt to ameliorate this limitation, we included additional coding variants to augment the coverage. Small sample size has always been a major limitation hindering the identification of rare variants, as demonstrated in this study. The sample size of the discovery stage was relatively small and therefore lacked statistical power to detect variants with modest effect size or very low frequency. Future large-scale meta-analysis with other Asian cohorts may serve to identify more functional variants that are specific to our population. Trans-ethnic meta-analysis will help to enhance the fine-mapping resolution of causal variants. An additional limitation of the present study may be the strategy used for selecting SNPs for replication.

In summary, the significant association of an Asian-specific coding variant, rs2233580 (p.Arg192His), with type 2 diabetes was identified in an exome-chip association analysis in a Chinese population. Our findings provide compelling evidence that *PAX4* could be a possible effector gene of the 7q32 locus and support its involvement in the pathogenesis of type 2 diabetes.

Acknowledgements The authors thank all the study participants and the clinical and research staff of the University of Hong Kong Theme-based Research Scheme (HKU-TRS) and the Guangzhou Biobank Cohort Study (GBCS) for their contribution to this research study.

Data availability Summary statistics of the current study are available from the corresponding author (ksllam@hku.hk) on reasonable request.

Funding This work was supported by: the Hong Kong Research Grant Council Theme Based Research Scheme (T12-705/11) and Collaborative Research Fund (HKU2/CRF/12R); the University of Hong Kong Foundation for Educational Development and Research (SN/1f/HKUF-DC/C20400.28505200); the Guangzhou Public Health Bureau (201102A211004011); and the Guangzhou Science and Technology Bureau, Guangzhou, People's Republic of China (2002Z2-E2051; 2012J5100041; 2013J4100031).

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

Contribution statement KSL, PCS, HFT and THL conceived the study, undertook project leadership and are guarantors of this work. CYYC and CST analysed the data and wrote the first draft of the manuscript. All authors contributed to the drafting and critical revision of the manuscript. PCS, CST and SSC interpreted the data and provided useful comments on the data analysis. CYYC, AX, CHL, KWA, LX, CHYF, KHK, WSC, YCW, MMAY, JSHH, YLJ, BMYC, KCBT, FZ, TZ, GNT, KKC and CQJ were involved in the sample collection, selection and phenotype data preparation for the University of Hong Kong Theme-based Research Scheme (HKU-TRS) and the Guangzhou Biobank Cohort Study (GBCS) cohorts. KSL, HFT, THL, GNT, KKC and CQJ were involved in the database management for the HKU-TRS and GBCS cohorts. All authors approved the final version of the manuscript.

References

- McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. *N Engl J Med* 363:2339–2350
- Tang CS, Zhang H, Cheung CY et al (2015) Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese. *Nat Commun* 6:10206
- Huyghe JR, Jackson AU, Fogarty MP et al (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 45:197–201
- Albrechtsen A, Grarup N, Li Y et al (2013) Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* 56:298–310
- Steinthorsdottir V, Thorleifsson G, Sulem P et al (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* 46:294–298
- Wessel J, Chu AY, Willems SM et al (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun* 6:5897
- Mahajan A, Sim X, Ng HJ et al (2015) Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. *PLoS Genet* 11:e1004876

8. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740–743
9. Nelson MR, Wegmann D, Ehm MG et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337:100–104
10. Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69
11. Abecasis GR, Auton A, Brooks LD et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
12. Mahajan A, Go MJ, Zhang W et al (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* 46:234–244
13. Zhang Y, Long J, Lu W et al (2014) Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies. *Cancer Epidemiol Biomarkers Prev* 23:622–628
14. Hui E, Yeung CY, Lee PC et al (2014) Elevated circulating pigment epithelium-derived factor predicts the progression of diabetic nephropathy in patients with type 2 diabetes. *J Clin Endocrinol Metab* 99:E2169–E2177
15. Chen C, Cheung BM, Tso AW et al (2011) High plasma level of fibroblast growth factor 21 is an Independent predictor of type 2 diabetes: a 5.4-year population-based prospective study in Chinese subjects. *Diabetes Care* 34:2113–2115
16. Jiang C, Thomas GN, Lam TH et al (2006) Cohort profile: the Guangzhou Biobank Cohort Study, a Guangzhou-Hong Kong-Birmingham collaboration. *Int J Epidemiol* 35:844–852
17. Jiang CQ, Lam TH, Lin JM et al (2010) An overview of the Guangzhou biobank cohort study-cardiovascular disease subcohort (GBCS-CVD): a platform for multidisciplinary collaboration. *J Hum Hypertens* 24:139–150
18. Guo Y, He J, Zhao S et al (2014) Illumina human exome genotyping array clustering and quality control. *Nat Protoc* 9:2643–2662
19. Shimajiri Y, Sanke T, Furuta H et al (2001) A missense mutation of Pax4 gene (R121W) is associated with type 2 diabetes in Japanese. *Diabetes* 50:2864–2869
20. Turner S, Armstrong LL, Bradford Y, et al (2011) Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 68:1.19.1–1.19.18
21. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res* 40:e53
22. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081
23. Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249
24. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
25. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26:2190–2191
26. Plengvidhya N, Kooptiwut S, Songtawee N et al (2007) PAX4 mutations in Thais with maturity onset diabetes of the young. *J Clin Endocrinol Metab* 92:2821–2826
27. Smith SB, Ee HC, Conners JR, German MS (1999) Paired-homeodomain transcription factor PAX4 acts as a transcriptional repressor in early pancreatic development. *Mol Cell Biol* 19:8272–8280
28. Bernardo AS, Hay CW, Docherty K (2008) Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic beta cell. *Mol Cell Endocrinol* 294:1–9
29. Blyszczuk P, Czyz J, Kania G et al (2003) Expression of Pax4 in embryonic stem cells promotes differentiation of nestin-positive progenitor and insulin-producing cells. *Proc Natl Acad Sci U S A* 100:998–1003
30. Sosa-Pineda B, Chowdhury K, Torres M, Oliver G, Gruss P (1997) The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. *Nature* 386:399–402
31. Petersen HV, Jorgensen MC, Andersen FG et al (2000) Pax4 represses pancreatic glucagon gene expression. *Mol Cell Biol Res Commun* 3:249–254
32. Cho YS, Chen CH, Hu C et al (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* 44:67–72
33. Ma RC, Hu C, Tam CH et al (2013) Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. *Diabetologia* 56:1291–1305
34. Kooptiwut S, Plengvidhya N, Chukijrungrat T et al (2012) Defective PAX4 R192H transcriptional repressor activities associated with maturity onset diabetes of the young and early onset-age of type 2 diabetes. *J Diabetes Complicat* 26:343–347
35. Xu W, Rould MA, Jun S, Desplan C, Pabo CO (1995) Crystal structure of a paired domain-DNA complex at 2.5 Å resolution reveals structural basis for Pax developmental mutations. *Cell* 80:639–650
36. Fuchsberger C, Flannick J, Teslovich TM et al (2016) The genetic architecture of type 2 diabetes. *Nature* 536:41–47
37. Voight BF, Scott LJ, Steinthorsdottir V et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579–589
38. Scott RA, Lagou V, Welch RP et al (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet* 44:991–1005
39. Steinthorsdottir V, Thorleifsson G, Reynisdottir I et al (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* 39:770–775
40. Liao JC, Yang TT, Weng RR, Kuo CT, Chang CW (2015) TTBK2: a tau protein kinase beyond tau phosphorylation. *Biomed Res Int* 2015:575170
41. Alesutan I, Sopjani M, Dermaku-Sopjani M, Munoz C, Voelkl J, Lang F (2012) Upregulation of Na-coupled glucose transporter SGLT1 by Tau tubulin kinase 2. *Cell Physiol Biochem* 30:458–465
42. Cariou B, Charbonnel B (2015) Sotagliflozin as a potential treatment for type 2 diabetes mellitus. *Expert Opin Investig Drugs* 24:1647–1656
43. Hart AW, Baeza N, Apelqvist A, Edlund H (2000) Attenuation of FGF signalling in mouse beta-cells leads to diabetes. *Nature* 408:864–868
44. Woo YC, Xu A, Wang Y, Lam KS (2013) Fibroblast growth factor 21 as an emerging metabolic regulator: clinical perspectives. *Clin Endocrinol* 78:489–496
45. Vordenbaumen S, Braukmann A, Petermann K et al (2011) Casein alpha s1 is expressed by human monocytes and upregulates the production of GM-CSF via p38 MAPK. *J Immunol* 186:592–601
46. Dinarello CA, Donath MY, Mandrup-Poulsen T (2010) Role of IL-1beta in type 2 diabetes. *Curr Opin Endocrinol Diabetes Obes* 17:314–321