Chapter 7

Investigation of *WFS1* common and rare variation for association with type 2 diabetes

7.1 Summary

Homozygous and compound heterozygous mutations in the Wolfram Syndrome gene 1 (*WFS1*) are associated with a rare syndrome including juvenile-onset non-autoimmune diabetes. In addition, it was recently discovered that risk of common type 2 diabetes is associated with common variants in *WFS1*, which map within a ~40kb linkage disequilibrium block on chromosome 4. In this study I attempted to refine the association signal by resequencing *WFS1* exons, splice junctions, UTR and putative regulatory regions in a subset of type 2 diabetes cases and disease-free controls, and performing an association study in three UK-based case-control studies. I also aimed to assess the contribution of rare (MAF<0.01) variation in *WFS1* to type 2 diabetes risk by deep resequencing of *WFS1* in 1235 cases and 1668 controls. These studies demonstrated association between type 2 diabetes and five previously untested *WFS1* SNPs, of which rs1046322 was the strongest (P = 0.008). However, due to high correlation between previously tested and untested SNPs it was difficult to refine the association signal to a smaller region. There was no statistical difference between incidence of type 2 diabetes in carriers and non-carriers of rare *WFS1* missense and nonsense changes. Nor was there a difference between carriers and non-carriers of rare synonymous changes, or rare variants with a high likelihood of having a functional effect on the protein. This suggested that rare variation in *WFS1* does not have large (OR>1.46) effect on risk of type 2 diabetes.

7.2 Introduction

7.2.1 WFS1 deficiency in humans and animal models

Wolfram syndrome (MIM 222300) is an autosomal recessive disorder characterised by diabetes insipidus, young onset non-autoimmune insulin-dependent diabetes mellitus, optic atrophy and deafness (Wolfram DJ 1938). Most patients carry loss-of-function mutations in the Wolfram syndrome gene 1 (*WFS1*), which encodes wolframin (Inoue et al. 1998; Strom et al. 1998). Over 100 mutations, including missense, nonsense and frameshift mutations, distributed throughout the gene have been described in Wolfram syndrome patients thus far (Cano et al. 2007), which appear to cause loss of function through depletion of wolframin rather than dysfunction of the protein (Hofmann and Bauer 2006).

Wfs1 knock-out mice (Ishihara et al. 2004) or mice with pancreatic ß-cell-specific deletion of Wfs1 (Riggs et al. 2005) show glucose intolerance and progressive pancreatic ß-cell loss. This phenotype appears to result from activation of ER stress pathways, impaired cell cycle progression, and enhanced apoptosis (Riggs et al. 2005; Yamada et al. 2006).

7.2.2 WFS1 has a role in ER calcium homeostasis and stress response

Wolframin is an endoplasmic reticulum (ER) membrane protein with nine transmembrane segments (Takeda et al. 2001). The C-terminal domain is located in the ER lumen, while the N-terminal domain extends into the cytoplasm (Hofmann et al. 2003). There is evidence that Wolframin functions as an ion channel or regulator of existing channels on the ER membrane (Osman et al. 2003) and that it positively modulates ER calcium uptake (Takei et al. 2006).

As described in Chapter 1, ER stress and the unfolded protein response (UPR) play a role in pancreatic ß-cell adaptation to the physiological demand for insulin and the

pathophysiology of insulin resistance, ß-cell failure and diabetes. All three ER stress pathways (PERK, IRE1, and ATF6) are activated by WFS1 deficiency in pancreatic ß-cells (Fonseca et al. 2005; Yamada et al. 2006). WFS1 is also transcriptionally up-regulated by ER stress inducing agents (Ueda et al. 2005), and contains a conserved sequence in its promoter region similar to the ER stress response element (ERSE) found in other components of the UPR (Kakiuchi et al. 2006; Ricketts et al. 2006). It is plausible, therefore, that WFS1 deficiency causes ß-cell apoptosis and glucose intolerance in mice and humans by triggering ER stress responses as a result of impaired ER calcium homeostasis and perturbing consequent cellular survival mechanisms such as the UPR.

7.2.3 Genetic variation in *WFS1* and type 2 diabetes (T2D)

As discussed in Chapter 6, I was involved in studies that demonstrated convincing association of common SNPs in *WFS1* with type 2 diabetes risk (Franks et al. 2008; Sandhu et al. 2007). However, there have been no attempts to refine the association signal or uncover the underlying functional variants. *WFS1* SNPs associated with type 2 diabetes were present in a block of high LD. The size of the interval between recombination hotspots flanking this block is ~ 68 Kb, defining a region in which the search for causal variants should start.

Another limitation of the association analyses conducted to date is that the common SNPs typed will not act as good proxies for rare variation in *WFS1*, and yet recent studies suggest that rare genetic variation with effects on complex traits that are intermediate between the effect size seen for common SNPs (OR<1.4) and fully penetrant Mendelian disease mutations (OR>2) can explain a substantial proportion of heritability (Bodmer and Bonilla 2008). Several publications from Hobbs and Cohen have reported enrichment of nonsynonymous mutations in candidate genes at one extreme of the population distribution of plasma lipoprotein traits (Cohen et al.

2004; Cohen et al. 2006; Romeo et al. 2007). Another study found an enrichment for rare nonsynonymous changes in monogenic obesity genes in obese compared to lean individuals (Ahituv et al. 2007). More recently, mutations in genes involved in renal salt handling were associated with lower blood pressure and protection from hypertension in a population-based cohort (Ji et al. 2008). There is also some evidence that intermediate frequency polymorphisms (MAF 0.01-0.05) contribute increased risk of disease compared to more common alleles (MAF>0.05). An intermediate frequency polymorphism (MAF ~ 0.02) in ANGPTL4 was associated with 10-15% lower triglyceride levels in population-based cohorts (Romeo et al. 2007). There is some anecdotal evidence that obligate carriers of Wolfram Syndrome mutations are more susceptible to T2D (Fraser and Gunn 1977). However, to my knowledge, there has been no systematic investigation of rare variation across the entire WFS1 gene for association with type 2 diabetes risk.

7.2.4 Aims

1. To attempt to refine the *WFS1* association signal by resequencing putative functional regions in a subset of type 2 diabetes case-control samples from the Sandhu *et al.* study and genotyping of newly discovered variants in additional case-control individuals.

2. To investigate whether rare variants within the coding sequence, splice sites, UTRs and conserved non-coding regions of *WFS1* contribute to type 2 diabetes risk.

7.3 Results and Discussion

7.3.1 Fine-mapping of *WFS1*

7.3.1.1 Identifying possible *WFS1* regulatory regions

Before sequencing *WFS1* in a subset of type 2 diabetes cases and controls, I looked for potential *WFS1* regulatory regions by identifying conserved sequences upstream of *WFS1* and within *WFS1* introns. Using different informatics software Sally Debenham (MRC Epidemiology Unit, Cambridge, UK) and I undertook multiple species alignments to look for evidence of conserved regions. Two regions in intron 1 appear to be well conserved: NCBI B36 coordinates 6325012-6325193 (181 bp) and 6325875-6326013 (138 bp), as well as two regions upstream of *WFS1*: NCBI B36 coordinates 6321756-6321858 (102 bp) and 6322195-6322297 (102 bp) (Figure 7.1).

7.3.1.2 SNP discovery

I sequenced *WFS1* exons, exon-intron boundaries, UTRs, and conserved upstream and intronic sequences in a subset of 96 Cambridgeshire case-control samples in order to detect known and novel *WFS1* SNPs (Table 7.1). These regions were considered more likely than non-coding non-conserved inter- and intra-genic sequence to harbour a true causative variant underlying the association with type 2 diabetes. Using this approach I identified 58 variants (Table 7.1), none of which mapped within conserved non-coding sequences. Nine SNPs altered the amino acid sequence, all but two of which (V333I and A559T) were predicted to have a damaging impact on protein function by at least one of SIFT, PolyPhen or PANTHER.

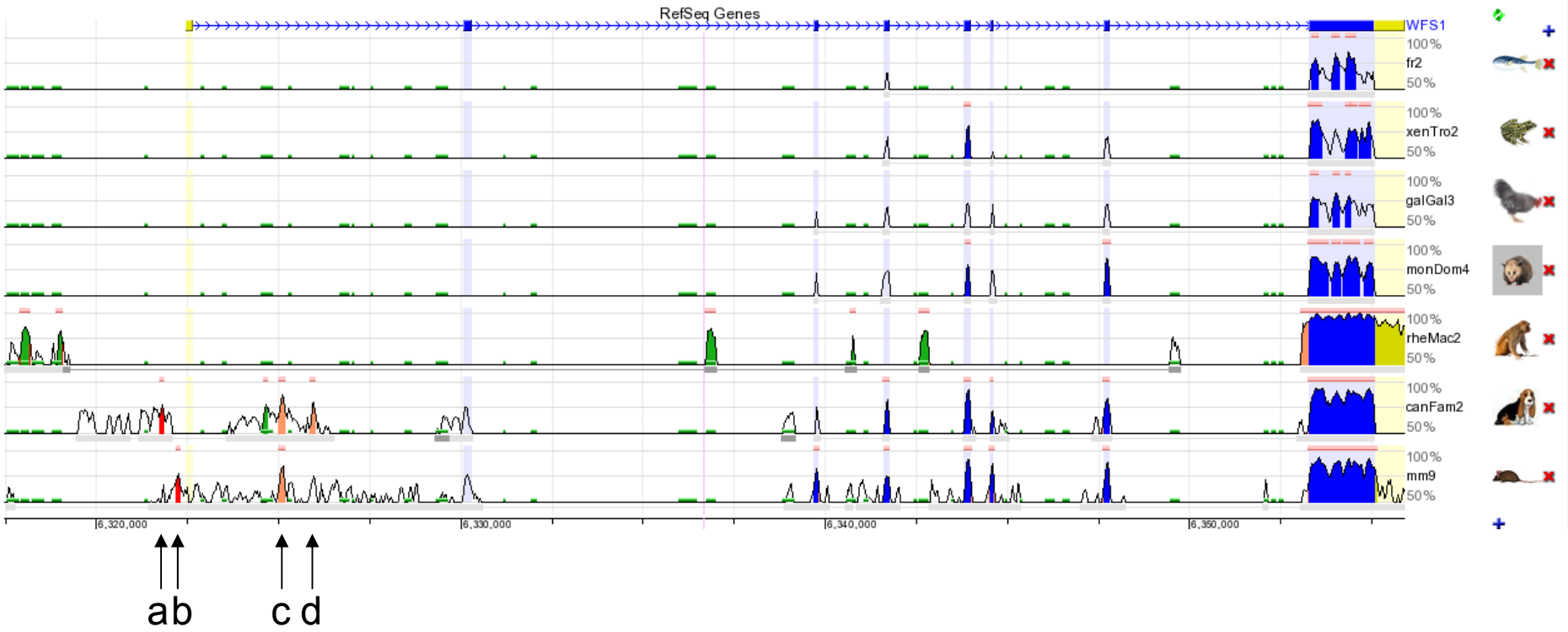

Figure 7.1 Evolutionary conserved regions (ECRs) in *WFS1* and 5000bp upstream
This figure was produced using the Dcode ECR browser (http://ecrbrowser.dcode.org/). Pink bars denote ECRs, blue bars denote exons and yellow bars denote UTR. The reference sequence is human and the graphs show sequence similarity to human in mouse, dog, monkey, opossum, chicken, frog, and fish (as shown on the right). Peak heights demonstrate the level of sequence similarity. White peaks indicate sequence with <80% similarity to human and <100bp in length, green peaks = transposons and simple repeats, blue peaks = exons, yellow peaks = UTR, salmon peaks = intron, and red peaks = intergenic sequence. Letters a and b = conserved upstream regions of 181 bp and 138 bp respectively, and c and d = conserved intronic regions of 102 bp.

Table 7.1 *WFS1* sequence variants detected in a subset of 96 Cambridgeshire case-control samples, with non-synonymous variants highlighted in blue

Genic position	Genomic position	Nucleotide substitution	Protein consequence	MAF in test samples	SNP ID
Upstream	6321944	T>A		0.33	rs4320200
Upstream	6321972	C>T		0.33	rs13107806
Upstream	6322051	G>C		0.34	rs13127445
Upstream	6322317	T>G		0.33	rs4273545
Intron 1	6324924	A>G		0.03	WFS1_1
Intron 1	6329948	T>C		0.20	rs10937714
Intron 2	6330405	A>G		0.50	rs28420833
Intron 3	6340039	A>T		0.02	WFS1_2
Intron 3	6341380	C>G		0.07	WFS1_3
Intron 3	6341421	G>A		0.01	WFS1_4
Intron 3	6341495	T>C		0.52	rs4688989
Intron 3	6341578	C>T		0.02	rs4688990
Intron 4	6341904	C>G		0.49	rs4689394
Intron 4	6343719	G>C		0.41	rs5018648
Intron 4	6343810	T>C		0.47	rs9998591
Intron 4	6343816	G>A		0.47	rs10010131
Exon 5	6343941	A>C	K193Q	0.01	WFS1_K193Q
Intron 5	6344138	G>C		0.47	rs9998835
Intron 5	6344253	C>T		0.47	rs10012946
Intron 5	6344351	C>T		0.47	rs13101355
Intron 5	6344378	G>A		0.48	rs13147655
Exon 6	6344597	G>C	R228R	0.30	rs7672995
Intron 6	6344703	T>C		0.12	rs7655482
Intron 6	6344739	G>A		0.40	rs11729672
Intron 6	6344746	G>A		0.01	WFS1_5
Intron 6	6344756	T>C		0.41	rs11725494
Intron 6	6344820	G>C		0.39	rs11725500
Intron 6	6344863	C>T		0.01	WFS1_6
Intron 6	6344868	A>G		0.39	rs4416547
Intron 6	6347348	G>T		0.07	rs12511742
Intron 6	6347438	G>A		0.02	WFS1_7
Exon 8	6353420	A>G	I333V	0.31	rs1801212
Exon 8	6353446	C>T	F341F	0.10	WFS1_F341F
Exon 8	6353608	T>C	V395V	0.45	rs1801206
Exon 8	6353717	C>G	L432V	0.01	rs35031397
Exon 8	6353731	C>T	T436T	0.01	WFS1_T436T
Exon 8	6353734	C>A	G437G	0.01	WFS1_G437G
Exon 8	6353790	G>A	R456H	0.04	rs1801208
Exon 8	6353923	T>C	N500N	0.42	rs1801214
Exon 8	6354098	G>A	A559T	0.01	WFS1_A559T
Exon 8	6354148	C>T	A575A	0.09	rs2230719
Exon 8	6354255	A>G	H611R	0.46	rs734312
Exon 8	6354475	G>A	A684A	0.01	WFS1_A684A
Exon 8	6354745	G>A	K774K	0.10	rs2230721
Exon 8	6354821	A>G	K800E	0.01	WFS1_K800E
Exon 8	6354856	G>A	K811K	0.48	rs1046314
Exon 8	6354875	C>T	R818C	0.01	rs35932623

Table 7.1 (continued). *WFS1* sequence variants detected in a subset of 96 Cambridgeshire case-control samples, with non-synonymous variants highlighted in blue

Genic position	Genomic position	Nucleotide substitution	Protein consequence	MAF in test samples	SNP ID
Exon 8	6354986	T>C	S855P	0.01	WFS1_S855P
Exon 8	6354988	A>G	S855S	0.36	rs1046316
Exon 8	6355034	G>A	V871M	0.02	WFS1_V871M
3'UTR	6355143	T>C		0.46	rs1046317
3'UTR	6355186	G>A		0.09	rs1802453
3'UTR	6355187	C>T		0.40	rs1046319
3'UTR	6355227	C>T		0.01	WFS1_8
3'UTR	6355245	G>A		0.50	rs1046320
3'UTR	6355349	G>A		0.12	rs1046322
3'UTR	6355370	A>G		0.03	rs1046325
3'UTR	6355384	C>T		0.01	WFS1_9

Genomic coordinates correspond to NCBI Build 36.

7.3.1.3 Selection of tagging SNPs

To assess whether any of the variants are highly correlated with the previously genotyped SNPs or whether any could act as proxies for one another, these 58 variants were uploaded into Haploview along with rs4689391, rs752854, rs6446482, and rs3821943 (which were genotyped in these samples as part of the original β-cell gene association study but were not covered by sequencing because they were outside the coding region and UTR). For reasons of power I decided to exclude rare SNPs (MAF<0.05) unless they altered the amino acid sequence, leaving 49 variants. Linkage disequilibrium between these 49 SNPs is indicated in Figure 7.2. Tagging SNPs were selected using an r^2 cut-off of 0.8, which generated 30 tagging SNPs. Seven of these had been genotyped as part of the original association study and so were removed from the selection (shown in red in Figure 7.2).

I then evaluated how well the selected tagging SNPs captured common variation in HapMap CEU trios within the linkage disequilibrium block containing the association signal (Rel 22/phase II April 2007) (Figure 7.3). This showed that 98% of common variants were captured and demonstrated that one additional SNP (rs12642481) was not well tagged. This SNP was force included into the tagging set (total number 24) to ensure all common variation was captured.

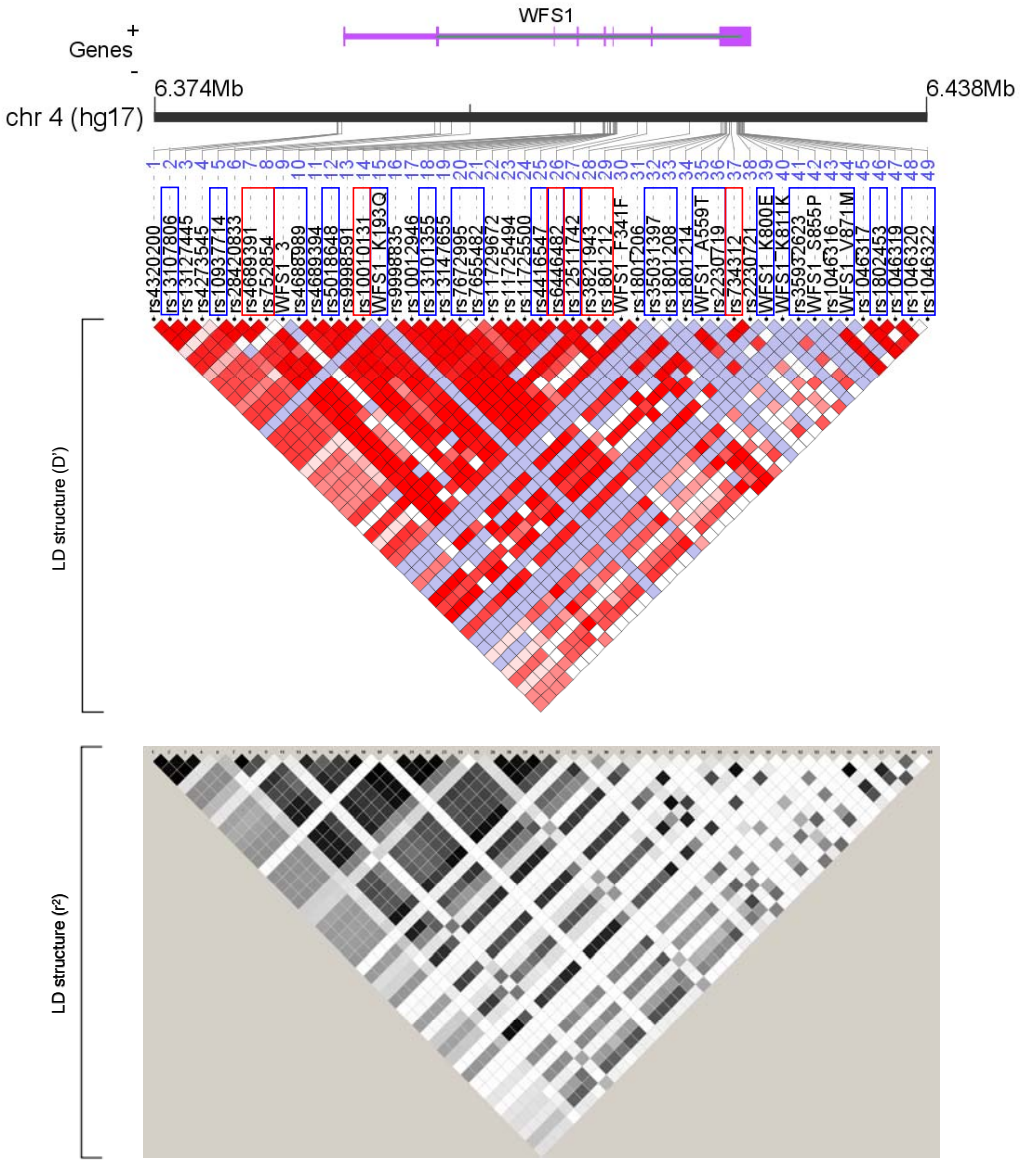

Figure 7.2 Feature map of the WFS1 gene showing SNPs discovered during resequencing and tagging SNPs

The positions of the 63 SNPs detected during sequencing of WFS1 in 96 Cambridgeshire cases and controls (including 4 additional SNPs genotyped during the original association study) are shown relative to the locus (purple) and chromosome 4 (black bar) (see text for details). The seven SNPs typed in the original studies are highlighted in red. Newly selected tagging SNPs are highlighted in blue. The bottom of the figure depicts two LD plots for the *WFS1* locus with pairwise LD values presented for SNPs. The upper plot presents LD as D' - see figure key for details. The figure was generated using LocusView (T. Petryshen, A. Kirby, M. Ainscow, unpublished software, available from the Broad Institute, Cambridge, MA (http://www.broad.mit.edu/mpg/locusview/). In the lower plot, LD among SNPs is given as r^2. r^2 values of 1.0 are represented by black diamonds, intermediate r^2 values are shown in grey and r^2 values of 0 as white. This plot was generated using Haploview (Barrett et al. 2005), available from the HapMap website (http://www.broad.mit.edu/mpg/haploview/index.php).

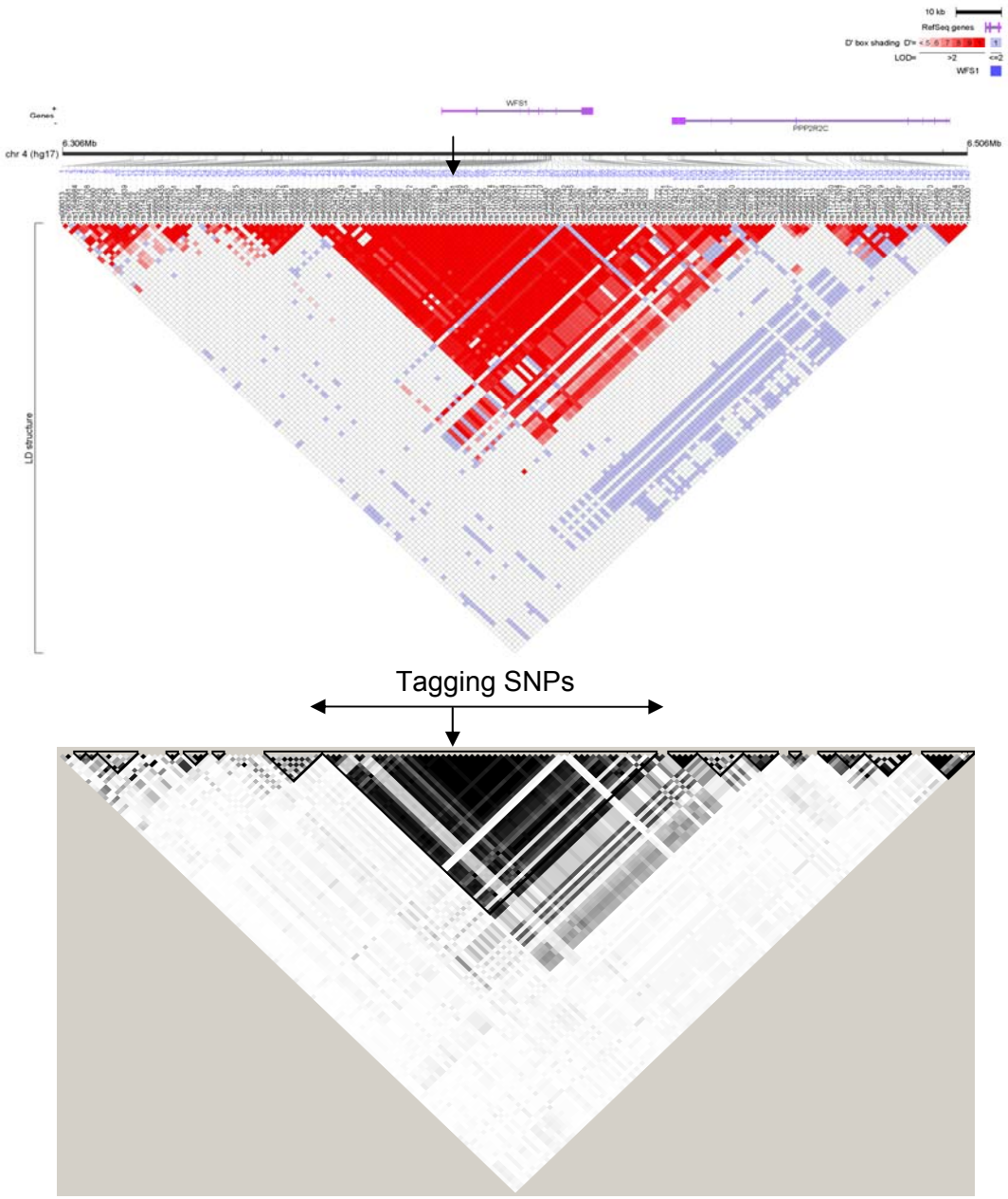

Figure 7.3 Feature map of the *WFS1* gene and flanking regions
The positions of SNPs genotyped in HapMap with a MAF≥ 0.05 are shown relative to known genes (purple) and chromosome 4 (black bar). The vertical arrows indicate the position of rs10010131, the SNP most significantly associated with risk of type 2 diabetes in the Sandhu *et al.* association study. The bottom of the figure depicts two LD plots for the *WFS1* locus with pairwise LD values presented for SNPs. The upper plot presents LD as D' - see figure key for details. The figure was generated using LocusView (T. Petryshen, A. Kirby, M. Ainscow, unpublished software, available from the Broad Institute, Cambridge, MA (http://www.broad.mit.edu/mpg/locusview/). In the lower plot, LD among SNPs is given as r^2. r^2 values of 1.0 are represented by black diamonds, intermediate r^2 values are shown in grey and r^2 values of 0 as white. This plot was generated using Haploview (Barrett et al. 2005), available from the HapMap website (http://www.broad.mit.edu/mpg/haploview/index.php). The horizontal arrows indicate the SNPs covered by my tagging SNP set.

7.3.1.4 Differences in call rate between cases and controls

Out of the 24 tagging SNPs selected for genotyping in Cambridgeshire, EPIC, and Exeter case-control studies, 21 passed clustering analysis (Table 7.2). Of these 21 SNPs, four had genotyping call rates that were statistically different between cases and controls. When each case-control study was analysed separately, it was clear that this disparity was driven in large part by Exeter genotyping, in which the call rate in cases was generally lower than in controls. Cases and controls were mixed on the plates so this cannot be due to differences in the genotyping quality between plates. The differentially called SNPs are spread throughout the gene, suggesting that a small regional duplication or deletion more common in Exeter cases and disrupting a primer/probe binding site is unlikely to account for all the observed discrepancies. The discrepancies were also across most DNA plates suggesting that it is unlikely to be a technical error concerning just a few DNA plates. In future studies, this could be confirmed by running the products of these PCR reactions in Exeter on gels to confirm success, and by sequencing across probe annealing sites. A large copy number variant (CNV) encompassing the whole region containing these SNPs could account for the lower call rates, however there are no annotated CNVs in the neighbourhood of *WFS1*. Another possible explanation for the differential call rates between cases and controls is a difference in the quality of the DNA, as cases and control samples were collected and extracted in separate studies. For this reason I decided to analyse only Cambridgeshire and EPIC studies (Cases = 854, Controls = 1242). A fresh supply of Exeter case samples may be required before repeating this genotyping.

Table 7.2 Quality control analyses of *WFS1* tagging SNP genotyping in UK case-control studies

						P value for the difference in call rate between cases and controls			
SNP	Genomic position	Protein consequence	MAF in all	HWE in controls	Call rate (All)	All	CCC	EPIC	Exeter
rs13107806	6321972		0.427	0.0403882	0.9335558	0.0009354	0.0855243	0.946636	0.000163
rs10937714	6329948		0.214	0.4157014	0.9262743	0.8371749	0.7272015	0.0409929	0.0048078
WFS1_3	6341380		0.05	0.5584238	0.9250607	0.0195131	0.0167841	0.8424461	0.0209893
rs4688989	6341495		0.399	0.1109792	0.9323422	0.0667898	0.016269	0.1618674	0.0009759
rs5018648	6343719		0.409	0.155748	0.9232403	0.1975619	0.0061325	0.2640687	0.017124
WFS1_K193Q	6343941	K193Q	0.005	*	0.964199	0.0001448	0.1067635	0.2083644	0.0013955
rs13101355	6344351		0.399	0.0495723	0.9123179	0.0000222	0.0046028	0.1793162	1.99E-09
rs7672995	6344597	R228R	0.317	0.2093614	0.9195995	0.0042431	0.5791705	0.8232458	1.09E-09
rs4416547	6344868		0.393	3.53E-06	0.8149272	0.019327	0.9772268	0.0370271	0.0976032
rs12511742	6347348		0.069	0.6957411	0.9535801	0.0095157	0.2939324	0.7612433	0.0004025
rs12642481	6351959		0.318	2.20E-07	0.8692355	0.3561777	0.1168826	0.6498275	0.0069747
rs35031397	6353717	L432V	0.004	*	0.9438714	0.6986788	0.6859223	0.0810481	0.0030927
rs1801208	6353790	R456H	0.049	0.0714386	0.9308252	0.2332388	0.3690197	0.0112531	0.0009366
WFS1_A559T	6354098	A559T	0.005	*	0.9611651	0.0004924	0.0681048	0.3805831	0.0015129
rs2230719	6354148	A575A	0.074	0.5240097	0.9414442	0.0775613	0.5540094	0.7915878	0.0056427
rs35932623	6354875	R818C	0.025	0.7604563	0.8658981	0.7940395	0.0352786	0.0136996	1.97E-14
WFS1_S855P	6354986	S855P	0.0003	*	0.8728762	0.0024014			
WFS1_V871M	6355034	V871M	0.024	1.12E-11	0.8834952	0.1891271	0.8149501	0.0295664	1.15E-06
rs1802453	6355186		0.092	0.9014183	0.9189927	0.9042388	0.7798427	0.2840468	0.0004025
rs1046320	6355245		0.414	0.4446617	0.8992718	0.3226095	0.9473739	0.7271306	0.0001631
rs1046322	6355349		0.122	0.9209425	0.9505461	0.2538266	0.0876172	0.0687599	0.0024334

Statistically significant *P* values are indicated in red. * = not applicable due to low frequency of minor allele (no rare homozygotes).

7.3.1.5 Association of *WFS1* SNPs with type 2 diabetes risk in Cambridgeshire and EPIC case-control studies

Of the 24 SNPs selected for genotyping 17 (71%) passed genotyping quality control in Cambridgeshire and EPIC and were taken forward for analysis, along with the seven SNPs genotyped as part of the original candidate gene association study. WFS1_S855P was only present in two individuals (MAF=0.0003), one case and one control, and was therefore excluded from further analysis. Given the high linkage disequilibrium across this region, the remaining 16 tagging SNPs that generated good quality genotypes (plus those genotyped as part of the original candidate gene study) captured 81% of the common (MAF>0.05) *WFS1* variation in the Cambridgeshire case-control samples used for SNP discovery. The 16 tagging SNPs covered 98% of the common *WFS1* variation in HapMap CEU trios, leaving only one intronic SNP (MAF = 0.24) untagged.

Eight SNPs were nominally associated with T2D risk (P<0.05) in a pooled analysis of Cambridgeshire and EPIC studies (Table 7.3). SNP rs10010131 is still the most statistically significant SNP of those seven genotyped as part of the original candidate gene association study (P = 0.024). However, four of my 16 new tagging SNPs show stronger association with T2D risk in Cambridgeshire and EPIC, rs1046320 being the most statistically significant (P = 0.008). SNP rs1046320 is in the 3'UTR of *WFS1* and is in high LD with the other nominally associated SNPs in this gene (Table 7.4). ClustalW mutliple sequence alignments showed that the nucleotide is only conserved in primates, not dog, cow, mouse or rat. Likelihood ratio tests demonstrated that adding SNP rs1046320 to logistic regression models containing one of the other seven statistically associated SNPs did not significantly improve the fit of these models (data not shown). Also, none of the statistically associated SNPs improved the fit of the simpler model containing only rs1046320. These SNPs are all correlated

in Cambridgeshire and EPIC samples (Table 7.4), indicating that they may all be linked to similar extents with the real causal allele(s) (which could be either untested or amongst them).

Table 7.3 Association of *WFS1* tagging SNPs with T2D risk in a pooled analysis of Cambridgeshire and EPIC case-control studies

SNP	Protein consequence	MAF	Odds ratio (95% CIs)	*P* odds ratio*
rs13107806	Conserved upstream	0.427	0.90 (0.79 - 1.02)	0.111
rs10937714	Intron 1	0.212	0.93 (0.79 - 1.09)	0.354
rs4689391	Intron 2	0.423	0.90 (0.79 - 1.03)	0.113
rs752854	Intron 2	0.344	0.87 (0.76 - 1.00)	***0.048***
WFS1_3	Intron 3	0.051	0.89 (0.66 - 1.21)	0.457
rs4688989	Intron 3	0.402	0.86 (0.75 - 0.98)	***0.025***
rs5018648	Intron 4	0.412	0.85 (0.74 - 0.97)	***0.014***
rs10010131	Intron 4	0.398	0.87 (0.76 - 0.98)	***0.024***
WFS1_K193Q	K193Q	0.004	1.00 (0.36 - 2.81)	0.997
rs13101355	Intron 5	0.4	0.85 (0.75 - 0.97)	***0.018***
rs7672995	R228R	0.316	0.84 (0.73 - 0.97)	***0.017***
rs6446482	Intron 6	0.405	0.87 (0.77 - 0.99)	***0.033***
rs12511742	Intron 6	0.072	0.93 (0.72 - 1.20)	0.584
rs3821943	Intron 7	0.457	0.91 (0.81 - 1.03)	0.146
rs1801212	I333V	0.28	0.90 (0.78 - 1.03)	0.137
rs35031397	L432V	0.004	1.10 (0.39 - 3.09)	0.856
rs1801208	R456H	0.046	1.25 (0.92 - 1.69)	0.152
WFS1_A559T	A559T	0.005	0.66 (0.25 - 1.75)	0.395
rs2230719	A575A	0.076	0.92 (0.72 - 1.18)	0.512
rs734312	H611R	0.455	0.93 (0.82 - 1.05)	0.247
rs1802453	3'UTR	0.089	0.93 (0.74 - 1.17)	0.539
rs1046320	3'UTR	0.419	0.83 (0.72 - 0.95)	***0.008***
rs1046322	3'UTR	0.119	1.01 (0.82 - 1.23)	0.948

* = the outcome of a logistic regression analysis. Bold text indicates significant *P*-values. Blue text highlights the most significantly associated SNP from the original study cohorts described in Chapter 6 (rs10010131) and the most significantly associated SNP in Cambridgeshire and EPIC case-control fine-mapping studies (rs1046320).

Table 7.4 Correlations among *WFS1* SNPs associated with T2D in the Cambridgeshire and EPIC case-control studies

	rs752854	rs4688989	rs5018648	**rs10010131**	rs13101355	rs7672995	rs6446482
rs752854							
rs4688989	0.71						
rs5018648	0.696	0.988					
rs10010131	0.717	0.963	0.967				
rs13101355	0.702	0.987	0.995	0.962			
rs7672995	0.59	0.7	0.699	0.686	0.7		
rs6446482	0.684	0.923	0.923	0.955	0.919	0.656	
rs1046320	0.655	0.932	0.939	0.92	0.939	0.666	0.883

LD values are r^2, where 1 denotes complete correlation and 0 denotes no correlation. Blue text highlights the SNPs from the original study described in Chapter 6. Bold text reveals the most significant SNPs in the original and fine-mapping studies.

7.3.1.6 Imputing untyped or failed SNPs

Using LD patterns between variants in the 96 sequenced Cambridgeshire case-control samples, I was able to impute genotypes of 25 additional variants detected during sequencing in all Cambridgeshire and EPIC samples. I also used LD patterns in HapMap CEU trios to impute HapMap SNPs in the interval between recombination hotspots flanking the association signal (Figure 7.4). SNP rs1046320 was still the most strongly associated SNP in Cambridgeshire and EPIC studies, except for one rare (MAF = 0.016) intronic SNP (rs7691824), imputed from HapMap (P = 0.0057). This SNP will need to be genotyped in Cambridgeshire and EPIC as imputation in this case is unlikely to be accurate considering the low frequency of the variant and its low correlation with typed SNPs.

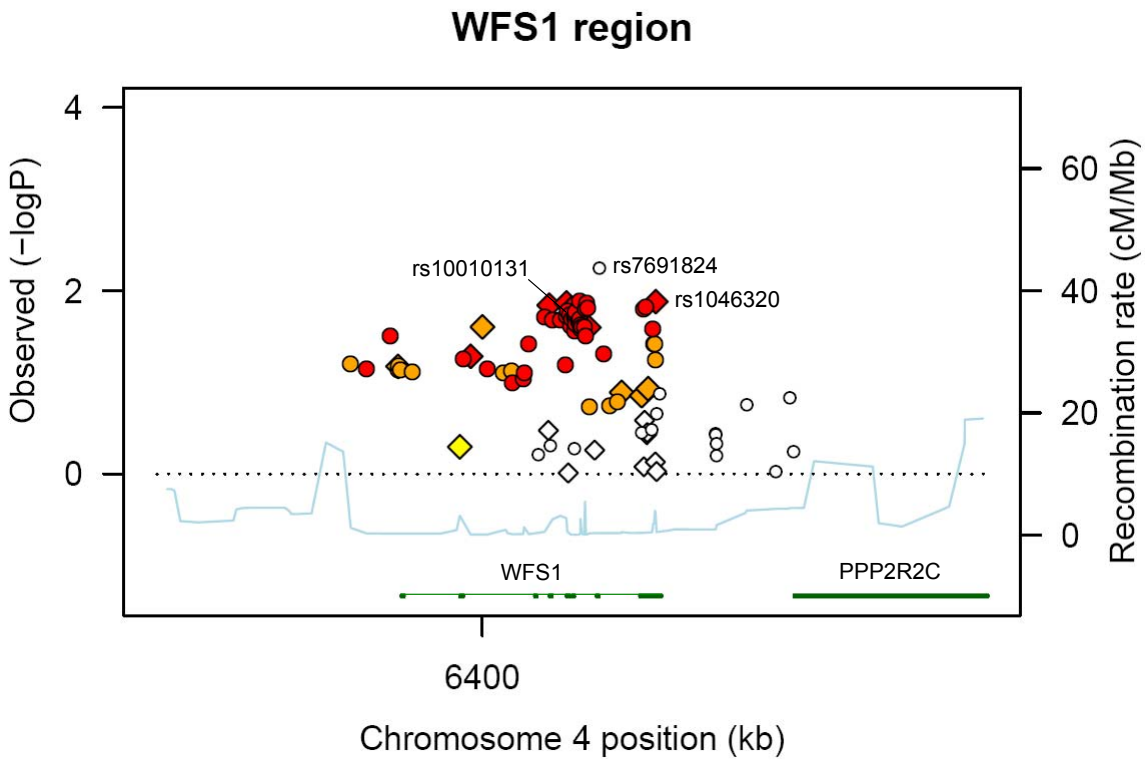

Figure 7.4 The statistical strength of the association of WFS1 tagging (diamonds) and imputed (circles) SNPs in the context of estimated recombination rates (blue line) and pairwise correlation between rs10010131 and surrounding markers
Red represents $r^2 > 0.85$, orange represents $0.5 < r^2 < 0.85$, yellow represents $0.2 < r^2 < 0.5$, and white represents $r^2 < 0.2$.

7.3.1.7 Combined analysis of rare variants in Cambridgeshire and EPIC

For very rare (MAF<0.005) non-synonymous variants WFS1_K193Q, rs35031397 (L432V), and WFS1_A559T, we had <80% power to detect effect sizes less than OR = 4.4, 4.5, and 3.7 respectively. Therefore, I tested whether the cumulative frequency of these variants influenced type 2 diabetes risk. In a combined analysis of WFS1_K193Q, rs35031397, and WFS1_A559T in Cambridgeshire and EPIC studies, there was no statistically significant difference in type 2 diabetes prevalence between carriers and wild-type individuals ($P = 0.709$).

7.3.1.8 Discussion

In this study I attempted to refine the previously reported association signal between *WFS1* variation and risk of T2D (Sandhu et al. 2007). I re-sequenced *WFS1* exons, splice junctions, UTR and putative regulatory regions in a subset of T2D cases (N=24) and controls (N=68) from the Cambridgeshire case-control study. I then selected tagging SNPs that covered common variation (MAF>0.05) and all non-synonymous variation detected by re-sequencing, as well as SNPs reported in HapMap CEU trios between recombination hotspots flanking the association signal. Tagging SNPs were genotyped and tested for association with T2D status in two UK case-control studies, Cambrideshire and EPIC case-control studies (854 cases and 1242 controls in total). Eight SNPs were nominally associated with T2D risk, five of which had not been tested in the Sandhu *et al.* study. Of these five previously untested SNPs, four showed stronger association with T2D than rs10010131. The strongest signal was from rs1046320 (P = 0.008). High correlation between these SNPs made it impossible to refine the association signal any further.

To test a denser set of variants across the region I imputed other variants discovered during sequencing of 96 Cambridgeshire case-control and other HapMap SNPs. Only one rare (MAF = 0.016) intronic SNP (rs7691824) imputed from HapMap showed stronger statistical association with T2D risk (P = 0.0057). This variant will need to be genotyped directly to confirm accurate imputation. The results from rs1046320 and rs7691824 need to be interpreted cautiously and repeated in other populations as, given that 89 variants were tested in the final analysis, the significance cut-off adjusted for multiple hypothesis testing using the Bonferroni correction is P = 0.000562.

This work demonstrates that while high linkage disequilibrium across regions of the genome is useful for minimising the amount of genotyping required to test the region for association with complex disease, it can compromise attempts to refine the association signal further. Discerning the underlying functional variants is particularly difficult when the surrounding variants are in nearly perfect linkage disequilibrium (r^2 > 0.9) because they give similar strengths of association. A more thorough approach could involve resequencing the entire interval between recombination hotspots (~68 Kb), rather than just those regions deemed most likely to harbour functional variation, and to analyse the sequence for copy number variations (CNVs) as well as SNPs. This would identify all possible genetic variants likely to impact disease risk. Also, as my search for conserved non-coding regions was restricted to *WFS1* intronic regions and 5 kb upstream and downstream of the gene, I will not have detected SNPs in potential regulatory regions towards the edges of the interval between recombination hotspots.

There is a risk that typing a more dense set of SNPs and CNVs may not add information due to high correlation between true functional variant(s) and other variants across the region. If this is so then studying populations with different and/or weaker patterns of linkage disequilibrium may help to refine the signal. For example, the LD block containing the *WFS1* gene is smaller in the HapMap samples of African descent, and correlation between SNPs is generally weaker (Figure 7.5). The LD between SNPs rs10010131 and rs1046320 is r^2 > 0.204 in YRI HapMap samples as opposed to r^2 > 0.92 in CEU samples. However, this study design carries certain caveats. The association signal in *WFS1* would need to be replicated in African populations, as the causal variants might not be present. Even if *WFS1* is a T2D susceptibility gene in Africans, the causal variant(s) (and those SNPs in LD with the causal variant(s)) may be different and would therefore be of limited value for refining the association in Europeans.

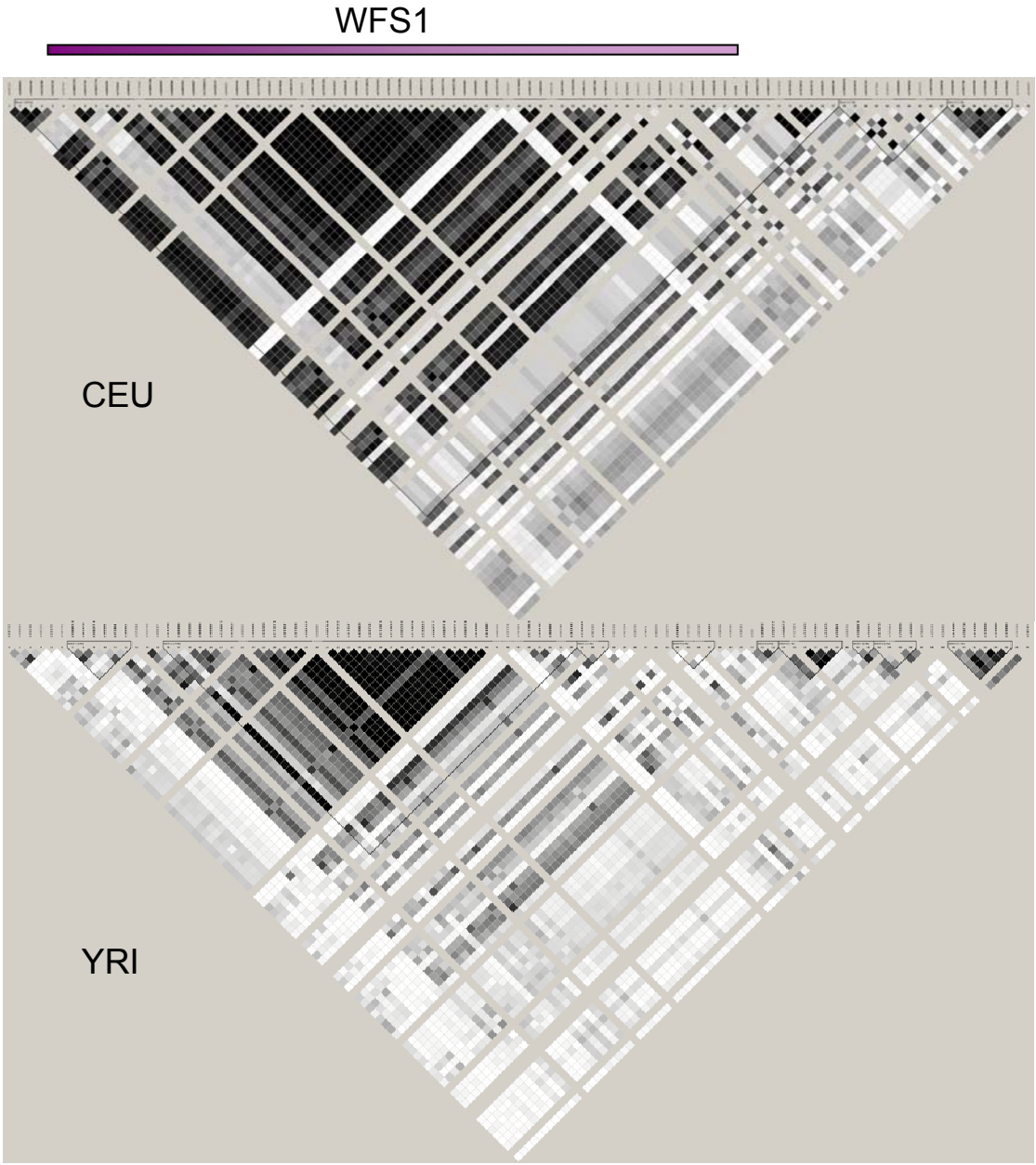

Figure 7.5 Patterns of linkage disequilibrium across the *WFS1* region in European (CEU) and African (YRI) samples
Presented are all SNPs in each population between NCBI build 36 coordinates 6315869 and 6379255. Gaps in the CEU and YRI LD plots represent SNPs not present in the respective samples. Linkage disequilibrium is measured by r2, with black diamonds representing high LD, white diamonds representing low LD, and grey diamonds representing intermediate levels of LD.

Another clue towards identifying true functional variants in *WFS1* would be the presence of eQTLs in the region – that is, genetic loci associated with changes in expression of *WFS1* or other genes in the region. However, no SNP or CNV within

the candidate interval has yet been found to be associated with gene expression variation in EBV-transformed lymphoblastoid cell lines from HapMap samples (GENEVAR http://www.sanger.ac.uk/humgen/genevar/).

Finally, haplotype analysis could be performed to test the joint actions of several SNPs across the *WFS1* region. It has been suggested that haplotype analyses may have better power than single SNP analyses to detect disease associations, as multiple SNPs in the haplotype may serve as better markers for the underlying risk allele(s). This approach might also help to focus resequencing efforts on individuals carrying a particular risk haplotype.

This study included several putative functional SNPs, including those that alter the amino acid sequence of Wolframin and those in highly conserved non-coding regions. However, none of the seven non-synonymous variants tested were associated with T2D risk in Cambridgeshire and EPIC studies. This could be because these variants do not affect risk of type 2 diabetes or the study could have been underpowered to detect their effect. I calculated that this study had <80% power to detect odds ratios less than 4.4, 4.5, 1.55 and 3.7 for SNPs WFS1_K193Q, rs35031397 (L432V), rs1801208 (R456H) and WFS1_A559T respectively. In a combined analysis of the very rarest non-synonymous SNPs (MAF≤0.005) I had <80% power to detect an odds ratio <2.55. Therefore, these variants cannot be ruled out as having a moderate impact on disease risk, but they are not causes of monogenic early-onset forms of diabetes as they were found in controls. Three SNPs found in an upstream conserved non-coding region by sequencing were tagged by rs13107806. However, this SNP was not statistically associated with T2D, indicating that upstream putative regulatory variants are not likely to contribute to risk of disease. The rs1046320 SNP is located in the 3'UTR and therefore could be affecting mRNA stability, processing and transport within the cell. Variants in the

3'UTR of genes have been found to impact disease. For example, a single nucleotide deletion in the 3'UTR of high mobility group A1 (*HMGA1*) gene reduces HMGA1 mRNA stability and expression and segregates with insulin resistance and type 2 diabetes in human subjects (Foti et al. 2005). More recently, it was suggested that SNPs in the 3'UTR of neurocalcin δ (*NCALD*) are associated decreased mRNA stability and risk of diabetic nephropathy (Kamiyama et al. 2007).

Four of the *WFS1* SNPs previously reported to be associated with T2D, rs4689391, rs3821943, rs1801212, and rs734312, did not reach statistical significance in this study. However, the direction and magnitude of their effects were similar. Given my sample size of 854 case-control pairs, I had between 28% and 35% power to detect an effect size OR 0.90 of SNPs with MAF between 0.28 and 0.48. Therefore, this study was statistically underpowered to detect the previously reported associations with these SNPs. Power could be improved by repeating the genotyping of the Exeter case-control study, as well as genotyping additional studies.

In conclusion, despite being statistically underpowered to detect the previously reported associations between *WFS1* SNPs and risk of T2D, I detected nominal associations with five previously untested SNPs, the strongest of which was rs1046320 (P = 0.008). Following imputation of HapMap SNPs, one imputed rare intronic SNP, rs7691824, was found to have even stronger association (P = 0.005). These SNPs will need to be genotyped in further case-control studies to investigate their impact on T2D risk.

7.3.2 Rare variant analysis

7.3.2.1 Resequencing of *WFS1*

I sequenced *WFS1* exons, exon-intron boundaries, UTRs, and conserved upstream and intronic sequences in the Cambridgeshire case-control study, the ADDITION study, and the MRC Ely cohort, which in total comprise 1668 controls and 1235 cases and 585 samples of unknown status (most of which were considered at high risk of developing type 2 diabetes). I detected 290 different sequence variants (Appendix Table A16) in these samples, 239 (82%) of which were novel. 235 (98%) of novel changes were rare (MAF<0.01) whereas only 15 (29%) of 51 known variants were rare, demonstrating the value of deep resequencing for identifying rare changes. 152 variants mapped within the coding region, of which 83 were non-synonymous, 66 were synonymous, and 3 were nonsense. There was a paucity of missense and nonsense changes with increasing minor-allele frequency, which is consistent with purifying selection acting on a significant fraction of such DNA sequence changes (Figure 7.6). Furthermore, there is an enrichment amongst low frequency variants for changes not detected in my sequencing of 96 Cambridgeshire case-control samples during the fine-mapping project (Figure 7.7). Though six variants were detected in non-coding regions with a high proportion of conserved residues, only two of these variants (in italics in Appendix Table A16) were actually conserved (the others falling between conserved nucleotides). Both were rare (MAF = 0.0003) and one was present only in cases and the other only in controls.

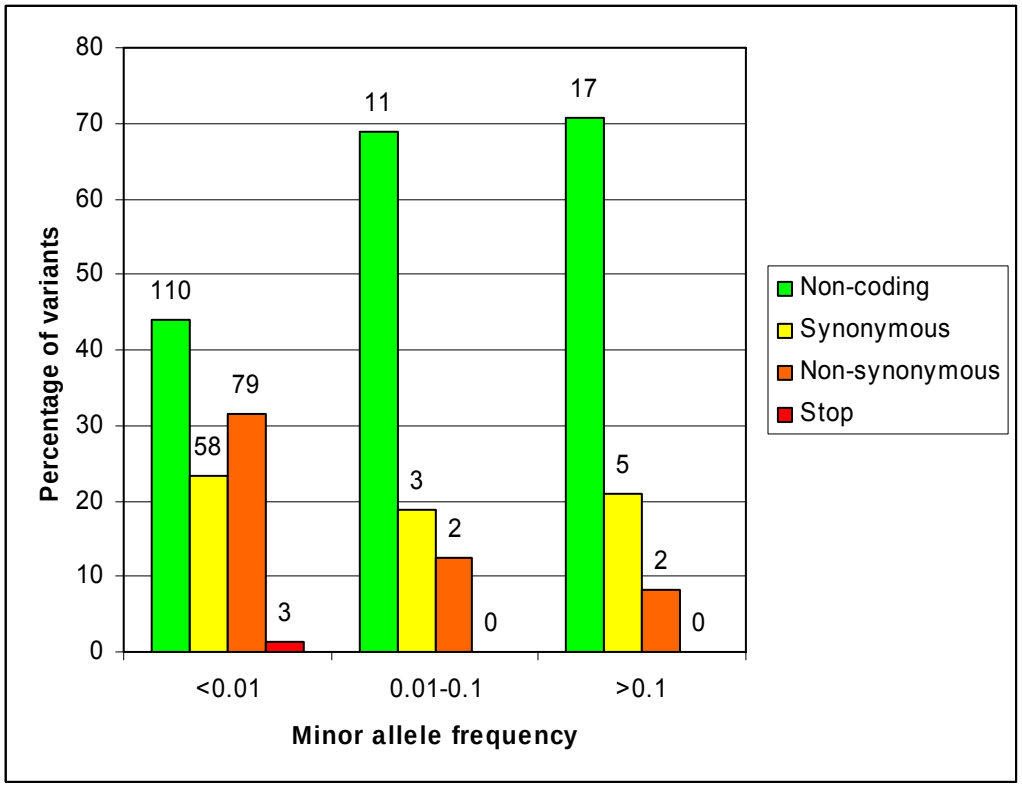

Figure 7.6 Distribution of types of *WFS1* variation discovered during resequencing of cases and controls at different minor allele frequency ranges
Numbers above the bars are the actual numbers found.

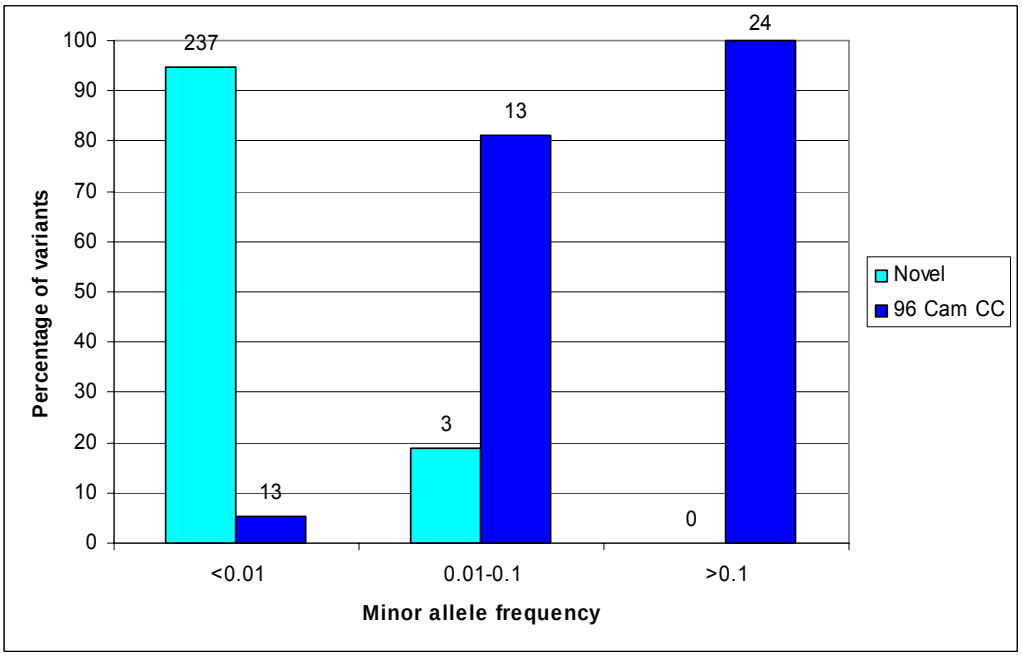

Figure 7.7 Distribution of novel and previously detected *WFS1* variation amongst different frequency ranges of changes discovered during resequencing of 1235 cases and 1668 controls
Actual numbers are shown above the bars.

7.3.2.2 Analysis of missense and nonsense variants with MAF<0.01

In my primary analysis I assessed the contribution of rare (MAF<0.01) missense and nonsense changes to risk of type 2 diabetes by comparing the odds of having type 2 diabetes in mutation carriers and non-carriers. A total of 82 missense and nonsense variants detected in our samples at a MAF<0.01 were included in this analysis (Appendix Table A17). Out of 2657 non-carriers, 1128 (42.45%) had type 2 diabetes, and out of 246 carriers, 107 (43.5%) had type 2 diabetes (Table 7.5). Therefore, there was no significant increase in risk of type 2 diabetes in carriers of rare missense and nonsense changes compared to non-carriers (OR = 1.04 (0.79-1.37), Fisher's exact $P = 0.788$).

Table 7.5 Number of cases and controls carrying missense or nonsense changes with MAF<0.01 vs wild-type

	Non-carriers	Carriers	Total
Controls	1,529	139	1,668
Cases	1,128	107	1,235
Total	2,657	246	2,903

OR = 1.04 (0.79-1.37), Fisher's exact $P = 0.788$.

As a small number of individuals carried more than one rare allele, I used logistic regression to assess the trend in the odds of disease with increasing number of mutations (Table 7.6). In this analysis, each additional mutation was associated with an extremely small and non-significant increase in risk of type 2 diabetes (OR = 1.01 ± 0.12, $P = 0.937$).

Table 7.6 Number of cases and controls carrying none, one, two, or three missense or nonsense changes with MAF<0.01

Number of mutations	0	1	2	3	Total
Controls	1,529	130	7	2	1,668
Cases	1,128	103	4	0	1,235
Total	2,657	233	11	2	2,903

OR = 1.01 ± 0.12, $P = 0.937$.

7.3.2.3 Analysis of synonymous variants with MAF<0.01

I decided to conduct a comparative study of synonymous variants, assumed to be functionally neutral, with MAF<0.01. This yielded similar results, though the effect sizes were larger. There was no significant change in odds of type 2 diabetes in carriers of at least one rare synonymous change compared to non-carriers (Table 7.7) (OR = 1.17 (0.81-1.68), P = 0.373), and no significant change in the odds per rare synonymous allele (Table 7.8) (OR = 1.32 ± 0.22, P = 0.089).

Table 7.7 Number of cases and controls carrying synonymous changes with MAF<0.01 vs wild-type

	Non-carriers	Carriers	Total
Controls	1,596	72	1,668
Cases	1,173	62	1,235
Total	2,769	134	2,903

OR = 1.17 (0.81-1.68), P = 0.373.

Table 7.8 Number of cases and controls carrying none, one, two, or three synonymous changes with MAF<0.01

Number of mutations	0	1	2	3	Total
Controls	1,596	72	0	0	1,668
Cases	1,173	55	6	1	1,235
Total	2,769	127	6	1	2,903

OR = 1.32 ± 0.22, P = 0.089.

7.3.2.4 Predicting variants with deleterious effects on the protein

Despite evidence that the majority of missense changes with MAF<0.01 have deleterious functional effects (Kryukov et al. 2007), I was concerned that I was diluting the effects of rare missense SNPs contributing to disease risk by analysing them with neutral missense changes. For this reason, I sought to identify non-synonymous and stop changes highly likely to impact on protein function and restrict the analysis to this group of variants (Table 7.9). I first looked for rare variants that had been shown biochemically to cause loss of function of Wolframin. R629W, W700X and P885L have all been shown to reduce the stability and half-life of wolframin (Hofmann and Bauer 2006) and have all been found in patients with

Wolfram Syndrome (Hardy et al. 1999; Hofmann and Bauer 2006; Kadayifci et al. 2001). Further variants with genetic evidence for involvement in Wolfram Syndrome include R558H (Colosimo et al. 2003), A559T and A671V (Smith et al. 2004), R708C (Tessa et al. 2001), E717K (Cryns et al. 2003), E776V (Smith et al. 2004) and R818C (Gomez-Zaera et al. 2001). Each missense variant was also entered into three different bioinformatics programs that predict functional impact based on sequence conservation and the biochemical properties of amino acids, SIFT, PolyPhen, and PANTHER. Furthermore, I carried out my own multiple sequence alignments to detect conservation of wild-type residues in monkey, mouse, rat, dog, chicken, frog, zebrafish, pufferfish and fruitfly. I noted that conservation in these multiple sequence alignments was not a good predictor of known inactivating *WFS1* mutations but biochemically proven mutations R629W and P885L were predicted damaging by all three bioinformatics programs. Therefore, I inferred 26 functionally important mutations based on a prediction of functional impact in SIFT, PolyPhen and PANTHER, and/or genetic/biochemical evidence for involvement in diabetes (Table 7.9).

Table 7.9 Known or inferred functional *WFS1* mutations

Chr:base	Variant	rs ID	Biochemical/genetic evidence	Pdel*	SIFT	PolyPhen	MAF in cases	MAF in controls	Conservation
4:6330207	R42X		Novel				0	0.0002998	
4:6343928	N188K		Novel	0.43033	affects protein	possibly damaging	0	0**	Low
4:6353402	L327F		Novel	0.55726	affects protein	possibly damaging	0.0004049	0	Vertebrate
4:6353502	C360Y		Novel	0.77597	affects protein	probably damaging	0.0008097	0	Vertebrate
4:6353739	F439C		Novel	0.59591	affects protein	probably damaging	0.0004049	0.0002998	Vertebrate
4:6353903	G494S		Novel	0.40523	affects protein	possibly damaging	0	0.0002998	Complete
4:6354096	R558H		Wolfram Syndrome (WS)	0.61604	affects protein	possibly damaging	0.0004049	0	Complete
4:6354098	A559T		WS and psychiatric disorders	0.34899	tolerated	benign	0.0048583	0.0029976	Low
4:6354105	I561S		Novel	0.44338	affects protein	possibly damaging	0	0.0005995	Low***
4:6354262	W613X		WS				0	0**	
4:6354306	T628M		Novel	0.74256	affects protein	possibly damaging	0.0004049	0	Vertebrate
4:6354308	R629W		WS & reduces half-life of wolframin	0.87775	affects protein	probably damaging	0.0004049	0	Low
4:6354435	A671V		WS and psychiatric disorders	0.21811	tolerated	benign	0.0012146	0	Low
4:6354443	G674R		Polymorphism	0.60981	affects protein	probably damaging	0.0004049	0	Low
4:6354449	R676C		Novel	0.75663	affects protein	probably damaging	0.0004049	0.0002998	Low
4:6354476	R685C		Polymorphism	0.84434	affects protein	probably damaging	0.0004049	0	Low***
4:6354522	W700X		WS & reduces half-life of wolframin				0	0**	
4:6354545	R708C		WS	0.77798	affects protein	probably damaging	0	0.0005995	Vertebrate
4:6354572	E717K		WS and psychiatric disorders	0.32653	tolerated	benign	0	0.0002998	Low
4:6354737	R772C		Psychiatric disorders	0.91098	affects protein	probably damaging	0	0.0008993	Low
4:6354750	E776V		WS	0.49302	affects protein	probably damaging	0.0040486	0.006295	Complete
4:6354792	S790W		Novel	0.71718	affects protein	possibly damaging	0	0.0002998	Low
4:6354875	R818C	rs35932623	WS and psychiatric disorders	0.68043	affects protein	possibly damaging	0.0048583	0.0053957	Low
4:6354917	R832C		Novel	0.75614	affects protein	probably damaging	0.0004049	0.0002998	Low***
4:6355061	D880N		Novel	0.49211	affects protein	possibly damaging	0	0.0002998	Vertebrate
4:6355077	P885L		WS & reduces half-life of wolframin	0.54691	affects protein	probably damaging	0	0.0002998	Complete

* Pdel score from PANTHER indicates the probability that an amino acid substitution will cause a deleterious effect on protein function based on alignment of evolutionarily related sequences (PANTHER classifies Pdel>0.38 as possibly deleterious) (continues on next page).

Table 7.9 legend continued.
** MAF = 0 in cases and controls shows that this variant was only found in samples of unknown disease status. N188K and W613X were found in ADDITION samples considered to be at high risk of developing diabetes (see Methods section), and W700X was detected in an Ely sample whose type 2 diabetes status was not recorded but quantitative trait data showed they had normal fasting glucose.
*** Amino acid with similar biochemical properties were conserved through evolution suggesting this locus may be of functional importance. Low = not well conserved. Vertebrate = conserved in all vertebrates. Complete = conserved in all species tested (monkey, mouse, rat, dog, chicken, frog, zebrafish, pufferfish and fruitfly).

7.3.2.5 Analysis of inferred functional variants with MAF<0.01

Out of 109 carriers of inferred functional mutations, 46 (42.2%) had type 2 diabetes, compared to 1189 cases (42.6%) in 2794 non-carriers (Table 7.10). This difference was not significant (OR = 0.99 (0.65-1.48), *P* = 1.00).

Table 7.10 Number of cases and controls carrying known and inferred functional *WFS1* mutations changes with MAF<0.01 vs wild-type

	Non-carriers	Carriers	Total
Controls	1,605	63	1,668
Cases	1,189	46	1,235
Total	2,794	109	2,903

OR = 0.99 (0.65-1.48), *P* = 1.00.

The trend in type 2 diabetes risk also decreased with increasing numbers of inferred mutations (Table 7.11).

Table 7.11 Number of cases and controls carrying none, one, two, or three known and inferred functional *WFS1* mutations with MAF<0.01

Number of mutations	0	1	2	3	Total
Controls	1,605	62	0	1	1,668
Cases	1,189	44	2	0	1,235
Total	2,794	106	2	1	2,903

7.3.2.6 Assessing association between disease status and a continuous measure of functionality of mutations

I decided to carry out an exploratory analysis to assess differences in the load of rare nonsynonymous variants between cases and controls, with mutations weighted by how likely they are to have deleterious effects on protein function. Instead of assigning carriers of mutations a score of 1 (as before), I weighted their score based on the PANTHER pdeleterious score for the mutation(s) they were carrying. In other words,

their score was now the sum total of the pdeleterious scores of all the rare (MAF<0.01) non-synonymous alleles they were carrying. To analyse differences between cases and controls I used a two sample T-test to assess the different in mean scores between case and control individuals. However, there was no significant difference between cases and controls (P = 0.5926).

7.3.2.7 The impact of intermediate frequency nonsynonymous SNPs (MAF 0.01-0.1) on risk of type 2 diabetes

Two nonsynonymous SNPs, V871M and R456H, had MAFs of 0.013 and 0.042 respectively. I had detected both SNPs during the fine-mapping study (Chapter 7.2.1) by sequencing a subset of 96 Cambridgeshire samples, but V871M failed genotyping and could not be imputed and R456H was not significantly associated with type 2 diabetes. I tested these in single SNP analyses in a pooled analysis of Cambridgeshire, ADDITION and Ely studies to assess association with type 2 diabetes risk. Neither V871M nor R456H were significantly associated with disease status (P = 0.132 and P = 0.249 respectively). This analysis had >80% power to detect effect sizes >1.93 and 1.45 for SNPs V871M and R456H respectively.

7.3.2.8 Discussion

Homozygous and compound heterozygous loss-of-function mutations in *WFS1* cause a Mendelian form of diabetes (Inoue et al. 1998; Strom et al. 1998), Wolfram Syndrome, and there is anecdotal evidence to suggest that obligate carriers of Wolfram Syndrome mutations have increased risk of type 2 diabetes (Fraser and Gunn 1977). Furthermore, numerous case-control studies have demonstrated association between polymorphisms in *WFS1* and risk of common type 2 diabetes (Franks et al. 2008; Sandhu et al. 2007), but were underpowered to detect moderate effect sizes of rare variants. Through deep resequencing of *WFS1* coding and conserved sequences in individuals with (N = 1235) and without (N = 1668) type 2 diabetes, I discovered 82 rare variants (MAF<0.01) which alter the amino acid sequence of Wolframin in 246 individuals. However, cases of type 2 diabetes were not significantly enriched amongst these rare variant carriers compared to non-carriers (P = 0.661). Carriers of rare variants deemed most likely to have a deleterious functional effect on the protein (based on web-based prediction algorithms and prior evidence for loss-of-function effects on the protein and/or co-segregation with Wolfram Syndrome) also did not have increased incidence of type 2 diabetes (P = 0.661). Nor was there a significantly different distribution of synonymous changes between cases and controls, as expected since these are predicted neutral.

Given the proportion of carriers in this case-control study (~8%), we had >80% power to detect OR>1.43 and >50% power to detect OR>1.29 (Power and Sample Size Program) (Dupont and Plummer 1990). This study was therefore well powered to detect previously reported effect sizes for rare variants on complex traits (the average being OR = 3.74) (Bodmer and Bonilla 2008). The impact of rare variants on risk of type 2 diabetes might have been diluted by pooling them with neutral rare variants for the analysis. Restricting

the analysis to those variants most likely to be functional reduced the frequency of the exposure (carrier status) to ~4%, but still retained >80% power to detect OR>1.65.

My study was underpowered to detect more modest effects (akin to those detected for common SNPs on T2D) of rarer variants. The Power and Sample Size Program (Dupont and Plummer 1990) indicates that a sample size of >22,000 cases-control pairs would be needed to have >80% power to detect effect sizes of rare missense and nonsense variants as low as OR = 1.09, assuming a similar proportion of carriers (8.5%) in the larger cohort. Also, studies with 80% power to detect modest effects (OR = 1.1) of SNPs with MAF=0.01-0.05 will require sample sizes between ~18,000-40,000 case-control pairs. However, it could be argued that such variants will not have an important impact on complex disease at a population-wide level.

In conclusion, I found no statistical enrichment for type 2 diabetes cases amongst individuals carrying at least one rare missense and/or nonsense change in *WFS1* compared to non-carriers. Given that my study was powered to detect effect sizes of OR>1.4, rare variants in *WFS1* are not likely to have an important impact on diabetes risk in UK populations.

7.4 Materials and Methods

7.4.1 Description of cohorts

7.4.2 Multiple sequence alignments

Sally Debenham used MultiPIP-maker (http://pipmaker.bx.psu.edu/pipmaker/) and VISTA MLAGAN (http://lagan.stanford.edu/lagan_web/index.shtml) to create alignments of the human *WFS1* genomic sequence and 5kb flanking regions and six other species (chimp, macaque, dog, cow, mouse and rat) to indicate the regions of conserved sequence. The human sequence was used as the reference sequence and was repeat masked. I identified two well conserved upstream regions using the Dcode ECR browser (http://ecrbrowser.dcode.org/).

7.4.3 PCR and sequencing

PCR, purification and sequencing of *WFS1* exons, exon-intron junctions, and UTR was performed using the standard protocol (Chapter 2.3.2). See Appendix Table A18 for primers and conditions. Sequencing of 96 samples from the Cambridgeshire case-control study were analysed, as part of the fine-mapping project, using Mutation Surveyor. Sequencing in the whole of the Cambridgeshire case-control study, ADDITION and MRC Ely studies was analysed using Gap4 (Chapter 2.3.6).

7.4.4 Genotyping

All 24 tagging SNPs passed assay design for genotyping on the Sequenom iPlex platform (Chapter 2.3.7.1.2). Primers and probes are listed in Appendix Table A19.

7.4.5 Quality control

Of 24 tagging SNPs genotyped in Cambridgeshire, EPIC and Exeter samples, three - rs7655482, rs1046316, and WFS1_K800E - were failed during manual confirmation of the genotype clusters. All remaining SNPs were checked for deviation from Hardy-Weinberg equilibrium (P<0.001), low call rates (N<85%) and significant discrepancy in call rate between cases and controls (P<0.001). Three SNPs, rs4416547, rs12642481, and WFS1_V871M, were not in Hardy-Weinberg and were not analysed. Except for rs35932623, all remaining tagging SNPs passed quality control in Cambridgeshire and EPIC case-control studies (Table 7.12). rs35932623 failed mostly in controls in Cambridgeshire and EPIC, but not Exeter. All failed SNPs except WFS1_K800E and WFS1_V871M were imputed, and WFS1_V871M was tested in Cambridgeshire and ADDITION/Ely samples as part of the rare variant analysis.

Table 7.12 QC in Cambridgeshire and EPIC samples

SNP	MAF	HWE in controls	Call rate	*P* difference*
rs13107806	0.427	0.04	0.918	0.02
rs10937714	0.212	0.98	0.902	0.557
rs4689391	0.423	0.06	0.921	0.473
rs752854	0.344	0.44	0.92	0.561
WFS1_3	0.051	0.12	0.904	0.027
rs4688989	0.402	0.14	0.914	0.291
rs5018648	0.412	0.11	0.902	0.333
rs10010131	0.398	0.02	0.998	0.521
WFS1_K193Q	0.004		0.959	0.005
rs13101355	0.4	0.05	0.898	0.072
rs7672995	0.316	0.09	0.905	0.771
rs6446482	0.405	0.17	0.997	0.711
rs12511742	0.072	0.53	0.949	0.275
rs3821943	0.457	0.08	0.996	0.65
rs1801212	0.28	0.22	1	0.407
rs35031397	0.004		0.93	0.664
rs1801208	0.046	0.06	0.908	0.06
WFS1_A559T	0.005		0.954	0.008
rs2230719	0.076	0.39	0.929	0.323
rs734312	0.455	0.08	0.988	0.739
rs35932623	0.027	1	0.85	**0.00001**
rs1802453	0.089	0.99	0.896	0.447
rs1046320	0.419	0.49	0.872	0.979
rs1046322	0.119	0.7	0.936	0.662

* between call rates in cases and controls.

SNPs and samples with call rate<0.9 were excluded from analysis of deep resequencing data. This lead to the elimination from analysis of 2 synonymous, 1 missense, and 2 non-coding variants out of a total of 322 variants detected. All these lost variants were rare (MAF<0.001). As manual editing of every common SNP call in sequence traces from ~3500 samples would have been too time consuming, only rare SNP calls were manually confirmed in raw sequence traces. All analysed SNPs were also tested for deviation from Hardy-Weinberg equilibrium and for statistically significant differences in call rate between cases and controls.

7.4.6 Statistical analysis

Statistical analyses were conducted using Stata v8.2. Hardy-Weinberg was assessed using the χ^2 statistic (1 df). Logistic regression was used to assess the contribution of individual SNPs under a log additive model (1 df) to risk of type 2 diabetes in the fine-mapping study, and to assess the trend in odds of type 2 diabetes in individuals with 0, 1, 2, and 3 rare variants in the rare variant analysis. Log likelihood ratio tests were also used to assess whether statistically associated SNPs independently contributed to risk of type 2 diabetes, comparing the log likelihood of a nested model (2 df) with that of the full model (3 df). The nested model contained only one SNP and the study cohort, whereas the full model contained an addition SNP to test if it contributes to disease independently of the variables in the nested model. For pooled analyses of Cambridgeshire and EPIC studies, and of Cambridgeshire, ADDITION and Ely studies, logistic regression with study as categorical covariate was carried out. The difference in odds of type 2 diabetes in carriers of rare variants vs non-carriers in the rare variant analysis was performed using Fisher's exact.

In all studies, linkage disequilibrium (LD), expressed as r^2, was calculated using Haploview v4.0 (http://www.broad.mit.edu/mpg/haploview) and power calculations were performed using Quanto v1.1.1 (http://hydra.usc.edu/gxe) and, for the analysis of type 2 diabetes in rare variant carriers vs non-carriers, the Power and Sample Size Program (Dupont and Plummer 1990).

7.4.7 Imputation

Imputation was performed by Eleanor Wheeler (Metabolic Disease Group, Wellcome Trust Sanger Institute). The best guess genotypes of SNPs were imputed with Cambridgeshire and EPIC separately using BIMBAM software

(http://stephenslab.uchicago.edu/software.html). SNPs that failed QC were not used to impute untyped SNPs but were instead imputed themselves.