

Genome-wide association study of longitudinal urinary albumin excretion in patients with type 1 diabetes

Anna M Hutchinson¹, Wei-Min Chen², Suna Onengut-Gumuscu², Paul Benitez-Aguirre³, Fergus J Cameron⁴, Scott T Chiesa⁵, Jennifer J Couper⁶, Maria E Craig³, Neil R. Dalton⁷, Denis Daneman⁸, Elizabeth A Davis⁹, John E Deanfield⁵, Kim C Donaghue³, Timothy W Jones⁹, Farid H Mahmud⁸, Sally M Marshall¹⁰, Andrew Neil¹¹, Stephen S Rich², M. Loredana Marcovecchio^{12*}, Chris Wallace^{1,13,14,*}

* These authors contributed equally to this work

¹ MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK.

² Center for Public Health Genomics, School of Medicine, University of Virginia, Charlottesville, Virginia, United States.

³ Paediatrics & Child Health, Children's Hospital, Westmead, University of Sydney, Sydney, Australia.

⁴ The Royal Children's Hospital, Melbourne, Australia.

⁵ Institute of Cardiovascular Science, University College London, London, UK.

⁶ Department of Endocrinology and Diabetes, Women's and Children's Hospital, North Adelaide, Australia.

⁷ Evelina London Children's Hospital, Guy's and St Thomas' NHS Foundation Trust, London, UK.

⁸ Department of Paediatrics, The Hospital for Sick Children and University of Toronto, Toronto, Ontario, Canada.

⁹ Telethon Kids Institute, University of Western Australia, Perth, Australia.

¹⁰ Institute of Cellular Medicine (Diabetes), Faculty of Clinical Medical Sciences, Newcastle University, Newcastle upon Tyne, UK.

¹¹ Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK.

¹² Department of Paediatrics, University of Cambridge, Cambridge, UK

¹³ Cambridge Institute of Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, CB2 0AW, UK.

¹⁴ Department of Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, CB2 2QQ, UK.

Corresponding author: Dr ML Marcovecchio

Department of Paediatrics, University of Cambridge, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0QQ, UK.

Phone: 01223 768615

Email: mlm45@medschl.cam.ac.uk

Abstract

Identifying genetic determinants for longitudinal changes in albumin excretion in individuals with type 1 diabetes may help identify those that are predisposed to renal, retinal and cardiovascular complications. Most studies have focussed on genetic predisposition to diabetic kidney disease and used cross-sectional measurements of urinary albumin excretion, but with limited success. Here, we utilise the wealth of longitudinal data and bio-samples collected from cohorts of childhood-onset type 1 diabetes followed over the last 30 years to describe a novel trajectory phenotype quantifying urinary albumin excretion changes during childhood and adolescence. We conducted a genome-wide association study and fine-mapping analysis for albumin excretion in 1584 individuals, finding one signal for cross-sectional albumin excretion close to *GALNTL6* (rs150766792), which validated in a previous independent study, and a novel genome-wide significant signal for albumin excretion trajectory in the *CDH18* gene region (rs145715205). Our trajectory phenotype quantifies albumin progression and offers a complementary measure to an albumin excretion phenotype based on a single measurement (i.e. most recent data collection) or an average of repeated measurements in longitudinal studies. It can be used to identify genetic or other risk factors which predict better or worse prognosis, thus facilitating the development of new preventive and therapeutic approaches.

Introduction

Diabetic kidney disease (DKD), the leading cause of end-stage renal disease for people with type 1 diabetes, remains a major determinant of morbidity and cardiovascular mortality (1,2).

DKD is characterised by progressive increases in urinary albumin excretion and decline in renal function, which follow different patterns, and occur at variable rates, over the lifetime of individuals with diabetes (1).

Extensive evidence indicates that increases in urinary albumin excretion, even within the normal range, predict renal, retinal and cardiovascular disease risk in adults and youth with type 1 diabetes (3–5). Although hyperglycemia is an important modifiable risk factor for increased albumin excretion, some patients still develop DKD irrespective of glycemic control, suggesting that other factors, either environmental or genetic, are implicated in the pathogenesis of this complication (1,6).

The identification of genetic determinants for albumin excretion is clinically important as it would enable the identification of individuals with type 1 diabetes that are predisposed to vascular complications, and potentially guide the development of new preventive and therapeutic approaches. This is particularly relevant to adolescents with type 1 diabetes, who have a higher risk of complications than adults due to longer diabetes exposure and often have suboptimal glycemic control (7).

To date, genome-wide association studies (GWAS) have only identified a small number of genetic variants that associate with DKD, or more specifically, urinary albumin excretion in individuals with type 1 diabetes (8–14); few studies have included cohorts of youth with type 1 diabetes, and most of the findings have not been replicated (6). This could be due to a

combination of relatively small sample sizes, differences in the study populations along with discrepancies across studies in the phenotype definition, ranging from microalbuminuria to end-stage renal disease (6).

Most genetic association studies for albumin-related phenotypes are based on cross-sectional measurements of albumin excretion (9,12–14), but cross-sectional analyses do not account for within-sample variability in albumin measurements over time (15). This is particularly true for youth with type 1 diabetes, who often show early increases in albumin excretion a few years after the diagnosis of diabetes with progression towards abnormal levels particularly during puberty, and potential regression into the normal range at the end of puberty (7,16). An alternative approach would be to investigate longitudinal changes in albumin excretion at an early stage in the development of DKD as a continuous trait, which would require summarising the longitudinal data in a manner suitable for GWAS. This is a challenging task due to the variability of duration and frequency of follow-up in many longitudinal studies.

Utilising the wealth of longitudinal data and bio-samples collected from carefully characterised childhood-onset type 1 diabetes cohorts prospectively followed over the last 30 years, we aimed to derive a novel quantitative phenotype of urinary albumin excretion changes during childhood and adolescence, and investigate novel GWAS genetic associations on this quantitative phenotype.

Research Design and Methods

Study Population

The study population consisted of young people with type 1 diabetes recruited into three longitudinal studies: the Oxford Regional Prospective Study (ORPS) (7), the Nephropathy Family Study (NFS) (17) and the Adolescents Type 1 diabetes cardiorenal Intervention Trial (AddIT) (18) (Table 1). For all three studies, ethical approval was obtained from the local ethics committees. Parents of participants provided written informed consent, and study participants were asked to provide their assent, until they reached an age when they could consent.

Full details of these cohorts have been previously reported (4,7,19,20). Briefly, the ORPS cohort was established in 1986, and approximately 500 children and adolescents diagnosed with type 1 diabetes when 16 years or younger were recruited between 1986-1996 in a defined geographical area (the Oxford Health Authority) and followed for up to 20 years. Annual assessments included: collection of clinical data, blood samples for HbA1c, and three consecutive urinary samples for the measurement of albumin-creatinine ratios (ACR). At baseline DNA samples were collected from a subgroup of the cohort ($n = 306$).

The NFS cohort was established between 2000 and 2005, when approximately 1000 young people (10-18 years), who had developed type 1 diabetes before the age of 16 years, were recruited in different UK centres (East Anglia, Birmingham, Bristol and Oxford regions). This cohort has been followed-up with annual assessments of height, weight, BMI, blood pressure, HbA1c and collection of three consecutive early-morning urine specimens for determination of the ACR. At baseline samples for DNA extraction were also collected.

The AddIT cohort included 847 adolescents with type 1 diabetes recruited between the years 2009 and 2013, across three countries (UK, Canada and Australia). In brief, 443 with an ACR in the upper tertile of the normal range (high-ACR group) were recruited to a randomised placebo-controlled trial of ACE-inhibitors and statins, and 404 with an ACR in the lower or middle tertile (low-ACR group) to a parallel observational study. Both groups underwent similar baseline and follow up assessments for a median of 3.9 years, based on a standardised protocol (4,19). For the present study, the study population consisted of 535 adolescents, representing a subgroup of those recruited into the AddIT Trial and parallel Observational study, with available DNA samples.

	All	ORPS	NFS	AddIT
N (% males)	1584 (54.7)	306 (55.4)	743 (56.3)	535 (53.1)
Age at diagnosis (years)	8.2 ± 3.7	8.3 ± 3.9	9.1 ± 3.5	7.6 ± 3.5
Age at first visit (years)	12.9 ± 2.5	10.9 ± 3.3	13.9 ± 2.1	12.5 ± 1.4
Duration of follow up (years)	4.6 ± 4.0	10.4 ± 4.2	2.1 ± 2.1	4.9 ± 1.8
ACR (mg/mmol)	1.8 ± 4.7	2.3 ± 7.5	1.7 ± 3.6	1.4 ± 2.3
BMI (kg/m²)	22.4 ± 4.1	21.4 ± 4.2	22.4 ± 3.9	23.0 ± 4.0
HbA1c (%)	8.6 (2)	9.3 (2.3)	8.9 (2.3)	8.3 (1.7)
HbA1c (mmol/mol)	70.0 (21.8)	78.1 (25.1)	73.8 (25.1)	67.2 (18.6)
Systolic blood pressure (mmHg)	116.2 ± 14.9	108.7 ± 14.8	118.9 ± 14.2	120.6 ± 13.0
Diastolic blood pressure (mmHg)	68.9 ± 9.6	69.4 ± 11.0	68.6 ± 9.7	68.7 ± 8.5

Table 1: Baseline clinical characteristics for the subjects included in our analysis. Values are expressed as mean±SD for normally distributed continuous variables, median (IQR) for non-normally distributed variables and percentage for categorical variables.

Albumin Excretion

All cohorts were monitored with annual ACR assessments. Urine samples were assessed centrally in a single reference laboratory as previously described (4). We excluded data where less than two consecutive measurements from a single visit were available (224 visits removed from a total of 11,601 visits), or where the ACR measurements from the same visit were highly variable (*standard error of the mean* > 30 mg/mmol; 19 data points removed), as this may indicate measurement errors. We then calculated the geometric mean of the consecutive ACR measurements.

GWAS Genotyping, Quality Control, and Imputation

We received genotype data for subjects in the ORPS, NFS and AdDIT cohorts. Study participants were genotyped in non-overlapping groups that were not cohort specific, and so we refer to these genotyping groups as groups 1, 2 and 3. Quality control (QC) and imputation were performed separately within each genotyping group ([Supplementary Methods](#); Supplementary Figs 1-5).

Participants in groups 1 and 2 underwent genotyping and QC centrally at the University of Virginia, which has been described previously (11). Samples were genotyped on the HumanCore BeadChip array (Illumina, San Diego, CA), which contains approximately 250,000 genome-wide tag SNPs and 200,000 exome-focused variants. We included an additional QC step to remove 261 chromosome X SNPs that were labelled as heterozygous in

males but did not reside in the pseudoautosomal region of chromosome X (Supplementary Table 1). For samples in group 3 which had not been quality controlled, we filtered for low-quality variants (e.g. high missingness, low minor allele frequency and excessive deviation from Hardy–Weinberg equilibrium) and samples (e.g. high missingness, sex mismatch, evidence of relatedness and extreme heterozygosity) (Supplementary Table 2). For samples with replicate data available, we selected the replicate with the best SNP coverage.

We used the Michigan Imputation Server (21) to impute missing genotypes for samples in each genotyping group separately, using the Haplotype Reference Consortium (HRC) panel of 64,976 human haplotypes at 39,235,157 SNPs as the reference data (version r1.1 2016) (22). We specified that the samples were of European ancestry and that any variants with low imputation quality ($R^2 < 0.3$) should be discarded. Following imputation, we removed samples with extreme proportions of homozygous variants and removed any monomorphic SNPs. Across the three genotyping groups, the final imputed and quality controlled data consisted of 11,483,371 SNPs in 2205 samples, for which 1584 samples had phenotype data available and were used in our analysis.

Multiple Imputation

There was ~15% missingness for HbA1c, blood pressure and BMI measurements across the study samples, which we accommodated for using a multiple imputation approach (23). Briefly, the multiple imputation pipeline consists of three stages: (i) imputation phase (ii) analysis phase and (iii) pooling phase. We used the mice (“Multivariate Imputation by Chained Equations”) R package (24) to impute 5 complete data sets, using a multilevel modelling approach with subject ID as the class variable and 10 iterations for each imputation, which allowed for sufficient mixing of chains (Supplementary Fig. 6). By

specifying a correlation coefficient $r > 0.1$ between variables for prediction, we selected ACR, treatment allocation and cohort for the prediction of missing HbA1c values, we selected sex, age, duration of diabetes, HbA1c, treatment allocation and cohort for the prediction of missing BMI values, we selected sex, age, duration of diabetes, HbA1c, treatment allocation (for the AdDIT cohort) and cohort for the prediction of missing systolic blood pressure (SBP) values and we selected ACR, age, duration of diabetes, HbA1c, treatment allocation and cohort for the prediction of missing diastolic blood pressure (DBP) values. To ensure that the imputed values were on the same scale as the observed values (e.g. $HbA1c > 0$), we used the squeeze function in the mice R package to transform the imputed data values to be within the range of observed data values.

In the analysis phase of multiple imputation, the analysis is implemented separately for each imputed data set. That is, sample-specific phenotypes were derived from each of the 5 imputed data sets (see “Phenotype Definitions” subsection) and 5 separate GWAS were conducted using these phenotype values (see “GWAS Analysis” subsection), resulting in 5 sets of GWAS summary statistics. In the final pooling phase, Rubin’s rules (25) were used to pool the statistics, ultimately deriving a final set of pooled effect sizes, standard errors and p-values (Supplementary Methods).

Phenotype Definitions

To derive a single albumin excretion phenotype summarising multiple longitudinal ACR measurements (Supplementary Fig. 7), we first used a linear regression model and selected the covariates for inclusion based on previous analyses in these cohorts (7,20). Specifically, the geometric mean of consecutive ACR measurements were logarithmically transformed and

used as the dependent variable, whilst sex, duration of diabetes, age, HbA1c, treatment allocation, cohort, BMI, SBP and DBP were selected as covariates in the model. For study samples in ORPS and NFS (where no treatments were allocated) the treatment allocation group was set to the control group in the AddIT cohort. We extracted the residuals (observed - fitted values) from the model, which correspond to covariate-adjusted ACR measurements.

We first defined a phenotype based on cross-sectional albumin excretion measurements (9,12–14) by selecting the covariate-adjusted ACR measurement from the most recent visit for each subject to derive the “latest ACR phenotype”. Second, we averaged the covariate-adjusted ACR measurements for each subject to derive the “average ACR phenotype” (20).

To better utilise the longitudinal data values, we derived a novel phenotype that models the trajectory of covariate-adjusted ACR measurements over time. For this, we used Bayesian multilevel modelling to fit average slopes as the trajectories. We opted to use a multilevel modelling approach as this allowed us to borrow information across subjects, in particular to improve parameter estimation for subjects with few data points available (and thus an uncertain trajectory). Explicitly, we used the `brm` function within the `brms` R package (26) which utilises the Stan software to fit the following Bayesian linear multivariate multilevel model:

$$y_i \sim \alpha + \alpha_i + \beta d_i + \beta_i d_i$$

where y_i is the residual for subject i measured at d_i years since their diagnosis. Parameters α and β represent the fixed effects and have independent improper flat priors, whilst parameters α_i and β_i represent random effects and are sampled from a multivariate normal prior with zero mean and an unknown variance-covariance matrix. We used the default prior for this variance-covariance matrix (see <https://cran.r->

project.org/web/packages/brms/vignettes/brms_overview.pdf for more details) and default parameter values, but increased the iterations per chain from 2000 to 8000 in order to ensure sufficient mixing of chains. We extracted the mean posterior slope coefficient for each subject (comprising both the fixed and random coefficients; $\beta + \beta_i$), and used these subject-specific values as our “trajectory phenotypes”.

GWAS Analysis

For study samples with both genotype and phenotype data available ($N = 1584$), we used the Gemma software (27) to perform genetic association tests for each of the 5 imputed data sets for each of our 3 phenotypes (“latest ACR phenotype”, “average ACR phenotype” and “trajectory phenotype”). The genomic control lambda values ranged between 0.95 and 0.99.

Because ACR has a right-skewed distribution, there is a potential for exaggerated p-values. To ensure our inference was robust, we repeated the complete analysis of each phenotype 1000 times with permuted phenotypes, and of a standard normal phenotype. For each null analysis, we saved the minimum p-value, and used the relationship between the minimum p-values from permuted and standard normal phenotype analysis to interpolate a more conservative adjusted p-value, upon which we based our GWAS inference.

We used the conventional significance threshold of $p < 5 \times 10^{-8}$ to call genome-wide significant signals and $p < 1 \times 10^{-6}$ for suggestive signals. To identify index SNPs in associated regions we selected the SNP with the smallest p-value in each 200-kb region. To aid comparison across phenotypes, we report standardised estimated effect sizes (β) and standard errors (SEs) (standardised so that the variance of the phenotype values equal 1).

Fine-Mapping Association Signals

We used Bayesian fine-mapping to prioritise the likely causal variants in associated regions, assuming a single causal variant per-region (25, 26). To define loci for fine-mapping we included all variants 200-kb either side of the index-SNP, and used the `corrcoverage` R package (29) to derive posterior probabilities of causality (PPs) for each SNP in the region, and to generate 95% credible sets.

To ensure that the single causal variant assumption made by `corrcoverage` was robust, we also performed fine-mapping using SuSiE (27, 28), which allows for multiple causal variants, using reference LD data from European samples in the 1000 Genomes Project (32).

Replication Analysis

We used the Type 1 Diabetes Knowledge Portal (T1DKP; <https://t1d.hugeamp.org/>, accessed 10 Feb 2022) to scrutinise significant signals identified by our analysis. We used results from the PheWAS association analysis for studies conducted on samples of European ancestry. In practice, the relevant studies on T1DKP were for the SUMMIT consortium (30, 31). We recorded the sample size, effect size (odds ratio due to case-control GWAS) and p-value for the relevant SNPs.

Data and Resource Availability

GWAS summary statistics will be made publicly available on the GWAS Catalog. Code to perform the analysis is available via <https://github.com/annahutch/T1D-ACR> and <https://github.com/chr1swallace/acr-p-values>. Phenotype data is available upon request.

Results

Novel Trajectory Phenotype

We describe a novel “trajectory phenotype” which effectively models the slope of covariate-adjusted ACR measurements over time. For this, we use a multilevel modelling approach that allows information to be borrowed across individuals, thereby protecting against extreme phenotype values for individuals with few data points. In practice, if an individual has many data points and therefore a more complete trajectory, then the method attenuates the slope of the observed trajectory only slightly towards the mean slope (Figs. 1A, 1B). In contrast, if an individual has very few data points and therefore a more uncertain trajectory, then the method more heavily attenuates the slope towards the mean (Figs. 1C, 1D).

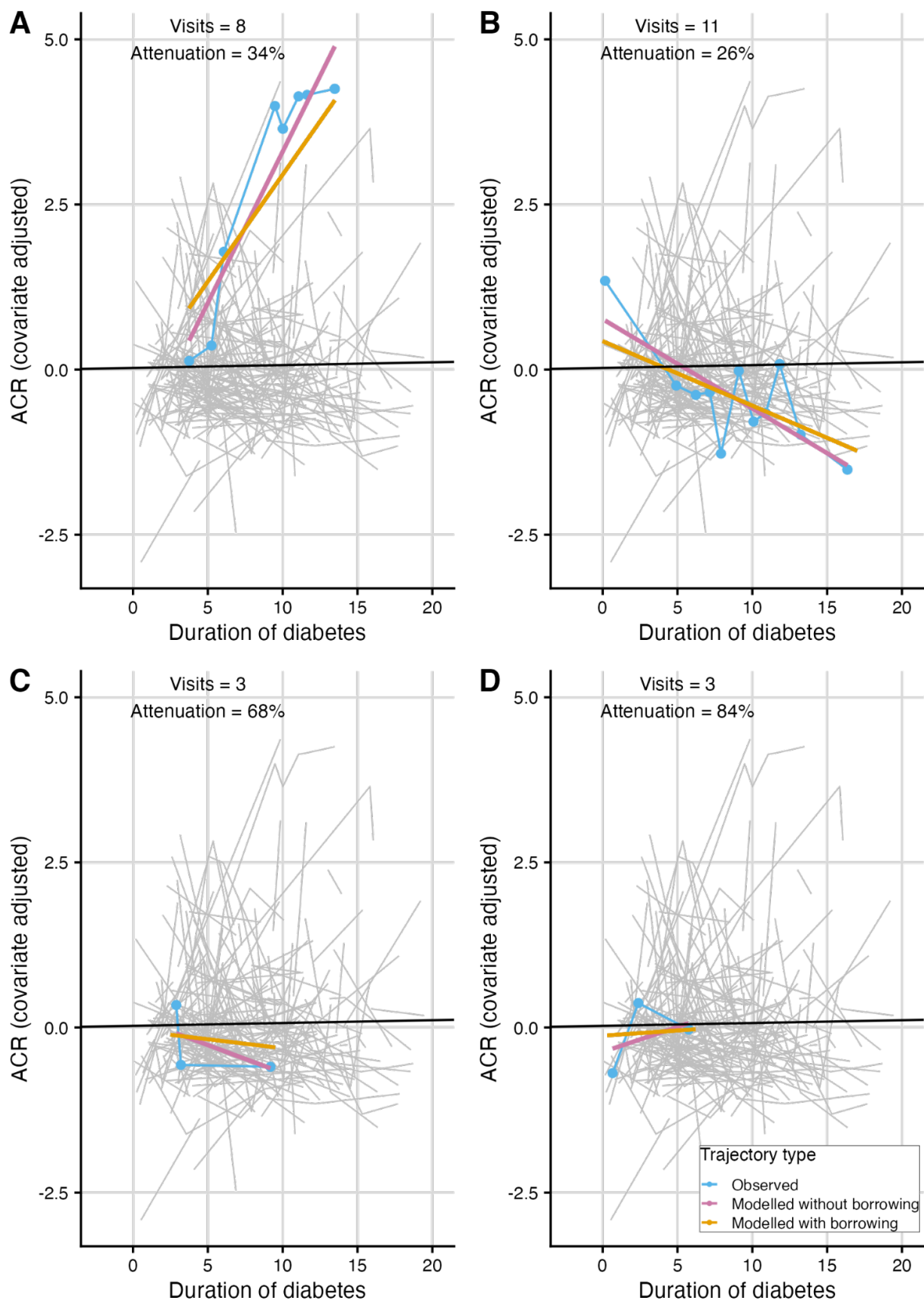


Fig. 1: Figure to illustrate modelling ACR trajectories to derive the trajectory phenotype. In plots of covariate-adjusted ACR measurements against duration of diabetes, we show the

observed trajectory (line between data points for each individual), the modelled trajectory without borrowing information across individuals (straight “best fit” line fitted to data points for each individual) and the modelled trajectory with borrowing across individuals (from which the slope equals the “trajectory phenotype”). Panels (A) and (B) highlight trajectories for individuals with many data points available, where the borrowing only attenuates the slope of the trajectory slightly. Panels (C) and (D) highlight trajectories for individuals with only three data points available, where the borrowing attenuates the slope of the trajectory more heavily. The grey lines represent a random sample of observed trajectories. The average trajectory across all subjects is shown in black. The number of visits and the attenuation between the trajectory modelled with borrowing and without borrowing is given in each panel. Attenuation was calculated as $1 - \text{slope of trajectory with borrowing} / \text{slope of trajectory without borrowing}$.

Comparison with Related Phenotypes

Previous studies of ACR have used cross-sectional ACR measurements (8, 11–13) or average ACR values (20). In our data, these values were highly correlated ($r = 0.77$). Our trajectory phenotype was also highly correlated with the latest ACR values ($r = 0.66$) and moderately correlated with the average ACR values ($r = 0.49$). This indicates that an individual’s trajectory is related to their average ACR, with more rapidly progressing trajectories corresponding to higher ACR values at the last visit.

Some understanding of the differences between these phenotypes may be revealed by considering individuals identified as having extreme phenotypic values. We found that the average ACR phenotype tended to identify those individuals with a single (or small number

of) extreme ACR measurements (Fig 2). In contrast, the trajectory phenotype highlights those individuals with strongly increasing (or decreasing) ACR measurements over time, thus monitoring albumin progression. The latest ACR phenotype only uses information from the most recent visit and is agnostic to any ACR measurements taken before this time.

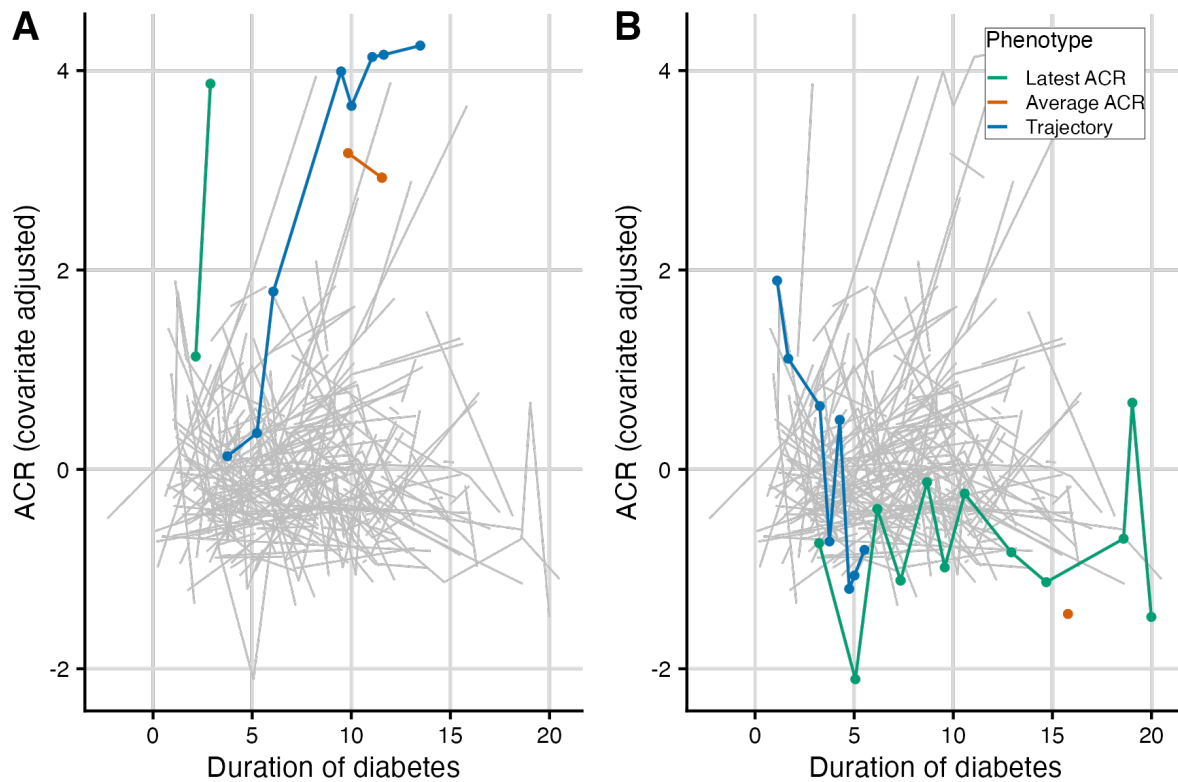


Fig. 2: Figure to illustrate observed trajectories for individuals with the highest and lowest phenotype values. In plots of covariate-adjusted ACR measurements against duration of diabetes, we highlight the observed trajectories for individuals with the (A) highest and (B) lowest phenotype value for each of the three phenotypes. The grey lines show a random sample of observed trajectories.

GWAS Results

We performed GWAS for all three ACR phenotypes (Fig. 3, Supplementary Fig. 8). At a suggestive significance threshold of $p < 1 \times 10^{-6}$, two signals were found to associate with the trajectory phenotype, three with the latest ACR phenotype (including one of the trajectory associated SNPs) and one with the average ACR phenotype (Table 2).

Of these suggestive signals, two exceeded the more stringent genome-wide significance threshold of $p < 5 \times 10^{-8}$. The first of these is for the latest ACR phenotype, and has index SNP rs150766792 for which the alternative allele (G) is relatively rare in European populations ($MAF = 0.0219$) (35). The SNP rs150766792 is closest to the long intergenic non-coding gene, *LINC02431* (27-kb from the start of the gene). The closest protein coding gene to rs150766792 is *GALNTL6* (800-kb to TSS).

The second genome-wide significant signal is for the trajectory phenotype and is driven by genetic variants residing in the first intron of the *CDH18* gene on chromosome 5. The alternative allele (C) at the index SNP, rs145715205, is relatively rare in European populations ($MAF = 0.0139$) (35). Interestingly, the latest ACR also showed suggestive association with rs145715205 in *CDH18* ($p = 6.4 \times 10^{-7}$) and also with rs79607490 ($P = 6.0 \times 10^{-7}$), which is upstream of *PCDH18* on chromosome 4, another member of the cadherin superfamily.

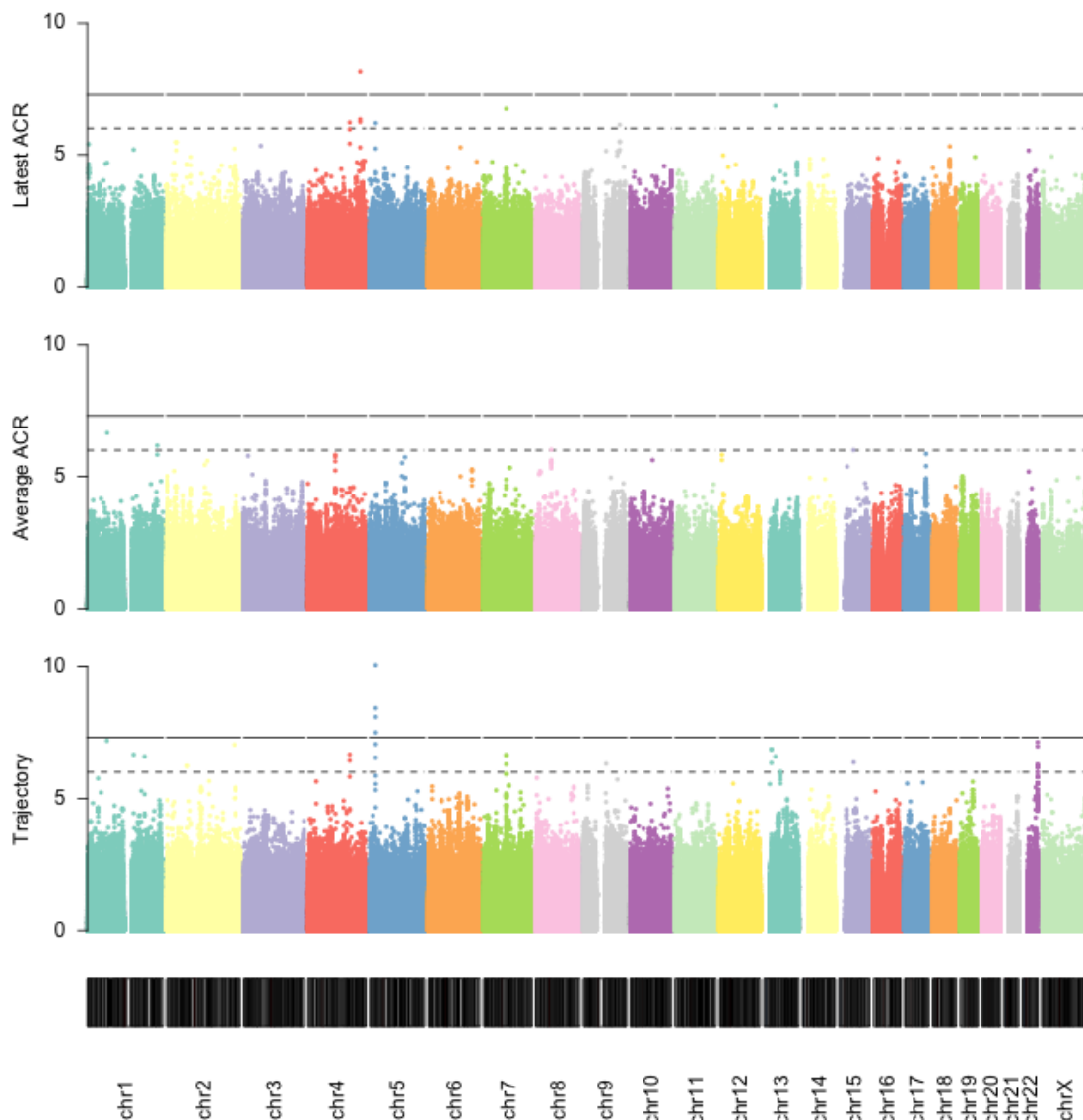


Fig. 3: Manhattan plots of GWAS results for (A) latest ACR phenotype (B) average ACR phenotype and (C) trajectory phenotype. Solid line on Manhattan plots show genome-wide significance threshold ($p = 5 \times 10^{-8}$) and dashed line shows suggestive significance threshold ($p = 1 \times 10^{-6}$). The y-axis is the $-\log_{10}$ transformed p-value and the x-axis is the genomic position.

rsID	Chr	bp	Alleles	Beta	SE	P raw	P adj	Closest gene
------	-----	----	---------	------	----	-------	-------	--------------

<i>Trajectory</i>								
rs115207889	1	64897735	T>G	0.929	0.171	6.53×10^{-8}	9.97×10^{-7}	<i>CACHD1</i>
rs145715205	5	20409042	A>C	0.883	0.135	8.9×10^{-11}	2.61×10^{-8}	<i>CDH18</i>
<i>Average ACR</i>								
rs112200619	1	65861210	T>C	0.743	0.141	2.20×10^{-7}	6.07×10^{-7}	<i>LEPROT</i>
<i>Latest ACR</i>								
rs150766792	4	171934268	A>G	0.729	0.125	6.94×10^{-9}	4.04×10^{-8}	<i>GALNTL6</i>
rs145715205	5	20409042	A>C	0.668	0.134	6.37×10^{-9}	1.66×10^{-8}	<i>CDH18</i>
rs187626285	7	73817796	G>A	0.828	0.158	1.81×10^{-7}	5.89×10^{-7}	<i>GTF2IRD1</i>
rs61947176	13	37885033	G>A	0.774	0.146	1.41×10^{-7}	4.85×10^{-7}	<i>CSNK1A1L</i>

Table 2: SNPs with suggestive evidence of association ($p < 1 \times 10^{-6}$) in any of the three GWAS phenotypes. For each index SNP, the table contains the rsID, the chromosome, the base position (bp, hg19), the alleles (reference > effect), the estimated effect size (beta), the standard error of the estimated effect size (SE), the observed and adjusted p-values and the closest protein coding gene. Note that rs145715205 appears twice, once for the trajectory and once for the latest phenotype.

Examining Significant Signals

We used Bayesian fine-mapping to prioritise the likely causal variants in the two genomic regions that reached genome-wide significance. In the *GALNTL6* gene region the index SNP, rs150766792, was found to be the most likely causal variant ($PP = 0.89$) and the 95%

credible set contained 2 additional variants (rs140921123 $PP = 0.040$; rs12502900 $PP = 0.054$) (Fig. 4, Supplementary Fig. 9).

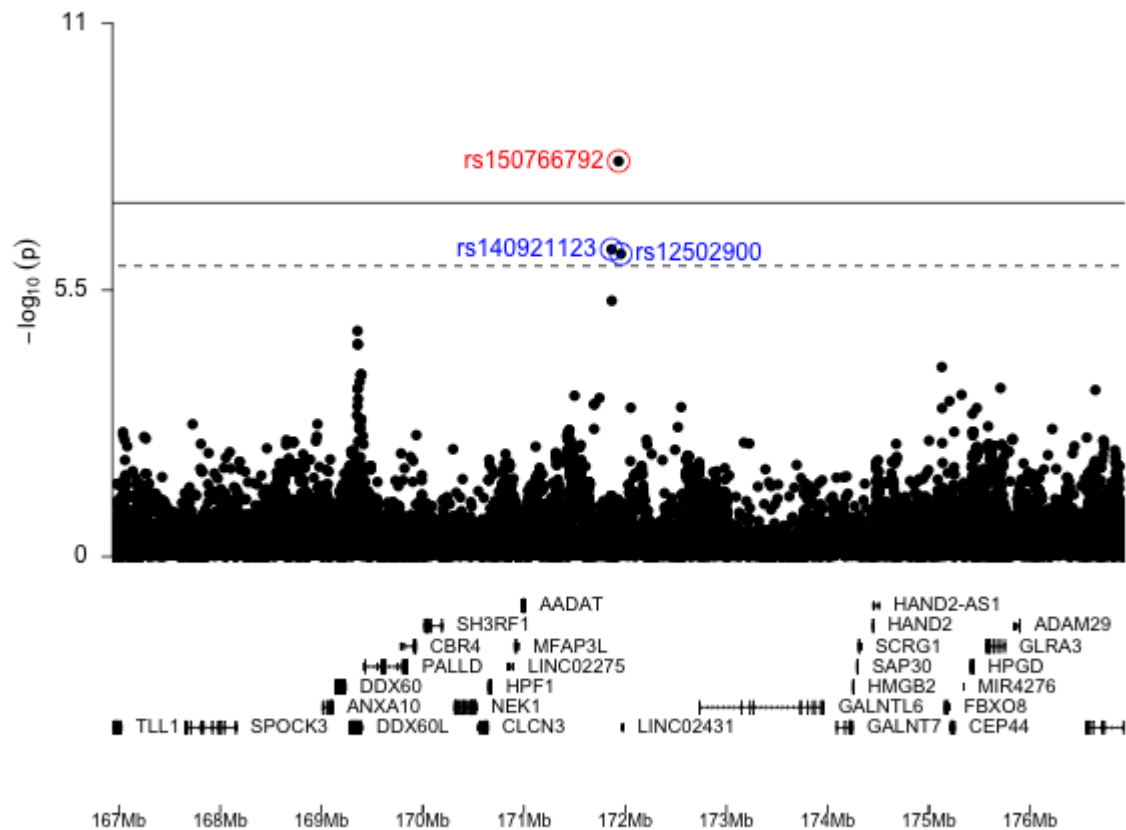


Fig. 4: Manhattan plot for the *GALNTL6* gene region (chr54:167000000-176000000). The index SNP in the region is highlighted and labelled in red. The two other SNPs included in the 95% credible set are highlighted and labelled in blue. Genes and genomic position (hg19) along chromosome 4 are shown in the bottom panel.

In a PheWAS association analysis on T1DKP, we found that the most strongly associated phenotypes with rs150766792 were microalbuminuria and all diabetic kidney disease, from the SUMMIT consortium (Table 3). Reassuringly, the effect sizes were directionally consistent with those from our GWAS, whereby the effect allele (G) increased risk. The

index SNP rs150766792 is correlated with the two other SNPs in the credible set ($r^2 = 0.4346$ with rs140921123 and $r^2 = 0.5695$ with rs12502900; estimated using 1000 Genomes Project European populations using the LDmatrix tool in LDlink (32,36)), and these SNPs also associated with microalbuminuria and all diabetic kidney disease in the SUMMIT consortium on T1DKP.

Study	Sample size	OR	P-value	Reference
Microalbumin in subjects with T1D or T2D	5,474	1.70	0.0003718	(30 ,31)
Microalbumin in subjects with T2D	3,464	1.90	0.001107	(34)
Microalbumin in subjects with T1D	1,747	1.41	0.1077	(33)
All diabetic kidney disease in subjects with T1D or T2D	9,307	1.50	0.001	(30 ,31)
All diabetic kidney disease in subjects with T2D	4,414	1.79	0.001436	(34)
All diabetic kidney disease in subjects with T1D	4,599	1.24	0.1554	(33)

Table 3: Replication results for rs150766792 from the Type 1 Diabetes Knowledge Portal (T1DKP; <https://t1d.hugeamp.org/>) in studies where subjects were of European ancestry. The effect allele for rs150766792 was G and the reference allele was A.

In the *CDH18* gene region the index SNP, rs145715205, was found to be the most likely causal variant ($PP = 0.975$) (Fig. 5). The PheWAS on T1DKP identified urinary sodium excretion as the most strongly associated phenotype with rs145715205 ($p = 0.007$) (37), with urinary albumin-to-creatinine ratio the second most strongly associated phenotype ($p = 0.019$) (35). SNP rs145715205 was not present in the SUMMIT data set so we were unable to check replication in this cohort.

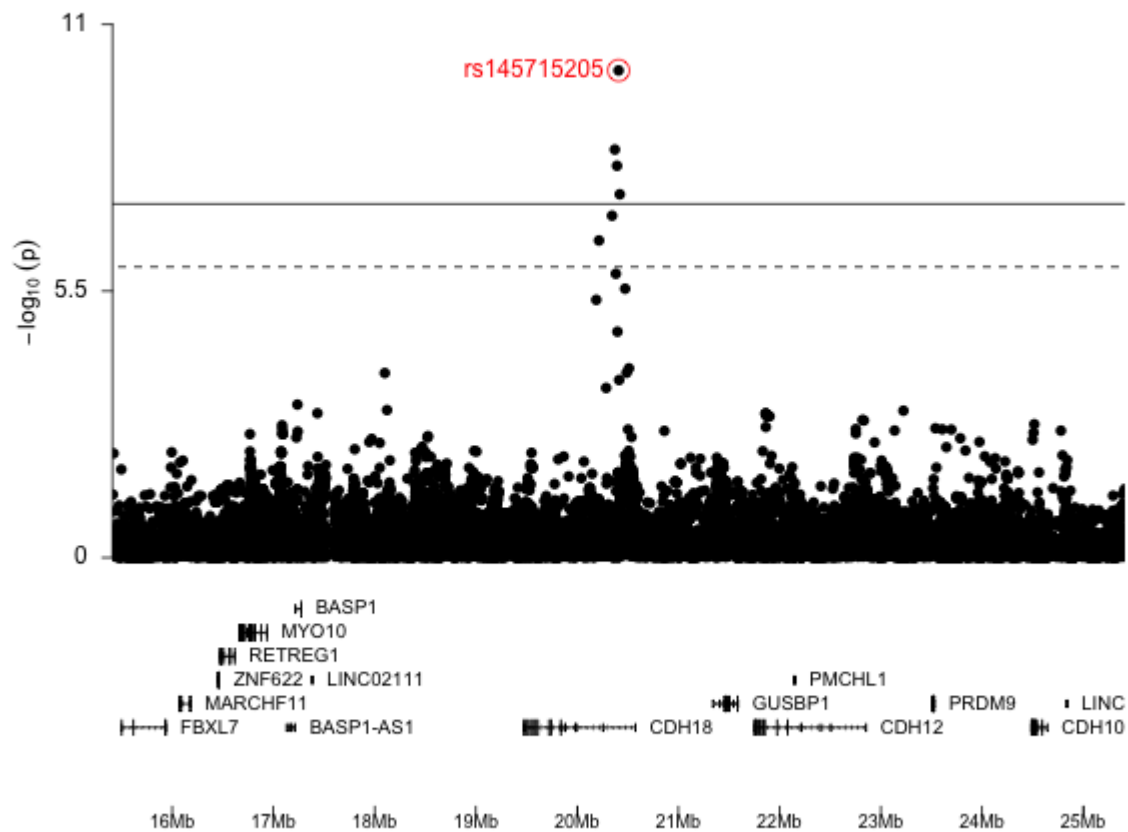


Fig. 5: Manhattan plot for the *CDH18* gene region (chr5:16000000-25000000). The index SNP in the region is highlighted and labelled in red. Genes and genomic position along chromosome 5 are shown in the bottom panel (hg19).

Discussion

We have described a novel albumin excretion phenotype, the trajectory phenotype, that is based on longitudinal data. This complements albumin-related phenotype definitions used in previous studies which are typically based on cross-sectional data by quantifying the slope, and not simply the central or final location of ACR measures. Estimated slopes will have a much higher degree of uncertainty for individuals with only a few measurements compared to those with more measurements. We accounted for this by taking a multilevel approach which naturally attenuates the most uncertain slopes towards the global mean, and reduces the risk of outlying observations having undue influence on GWAS estimates. Given the variability in numbers of observations per individual, we considered this approach was likely to be more robust than other approaches for summarising longitudinal data, such as latent trajectory modelling that groups individuals into trajectory groups whereby individuals within the same trajectory group have similar trajectories. Whilst such approaches have the advantage of being able to accommodate non-linear trajectories, this requires many more degrees of freedom, which in turn requires more observations, amplifying the difficulty assigning accurate phenotypes to individuals with fewer observations.

To the best of our knowledge, no other study has derived a phenotype that models longitudinal ACR changes for use as a continuous trait in GWAS. Whilst we used this phenotype to identify genetic variants which predict better or worse ACR trajectories, a similar approach could be considered for studying potential non-genetic predictors. Such approaches could help towards the development of a personalised prediction tool to help guide therapeutic decisions, for example commencing statin usage earlier if genetic or other risk factors are present.

Our signal close to *GALNTL6* (lead SNP rs150766792) for the latest ACR phenotype was replicated in the SUMMIT cohort, which was reassuring because our cross-sectional latest ACR phenotype more closely resembled the microalbuminuria and DKD case definitions in the SUMMIT GWAS than the other phenotypes. We identified a novel signal in the *CDH18* gene (lead SNP rs145715205) when using our trajectory phenotype. We were unable to replicate this signal, which may reflect that no previous studies have looked for genetic associations with albuminuria progression in type 1 diabetes patients.

The *CDH18* gene, located in the region of chromosome 5 encodes for cadherin 18, which belongs to the large cadherin superfamily, a class of calcium-dependent trans-membrane proteins. Cadherin 18 is expressed in various tissues but appears to be confined primarily to the central nervous system (38). SNPs within *CDH18* have been previously associated with gestational diabetes (39) and with metabolic syndrome-related traits, such as obesity, insulin resistance and blood pressure (40). Although there is no clear evidence yet for *CDH18* to be expressed in the kidney or vascular system, and there is no previous evidence of association with albumin excretions or more general DKD, the role in metabolic syndrome-related traits is of interest. Indeed, urinary albumin excretion and DKD are closely associated with features of the metabolic syndrome, such as obesity, insulin resistance and hypertension (41,42).

The *GALNTL6* gene encodes the membrane-bound protein N-acetylgalactosaminyltransferase like 6, which has a significant role in the pathway of protein glycosylation (43,44). There are no previous reports of a role of this gene in the context of diabetes and its vascular complications. However, a potential role could be speculated given that protein glycosylation is a well known process implicated in the pathogenesis of vascular complications of diabetes (45).

It will therefore be important to validate our findings, for example in larger study cohorts or cohorts of differing ancestry. Functional studies will also be required to investigate the functional roles of rs145715205 and rs150766792 in albuminuria and its progression in patients with type 1 diabetes. Of interest, this study also identified a new approach to assess longitudinal ACR data, which could be applied not only in future GWAS studies, but could also help towards the development of a personalised prediction tool to help guide therapeutic decisions, for example commencing more intensive drug interventions earlier if genetic or other risk factors are present.

Acknowledgments

The authors gratefully acknowledge the contribution made to this study by David Dunger who was the Chief Investigator of the ORPS/NFS and AddIT studies prior to his death in July 2021. We would like to thank Rany Salem for advice on genotype data cleaning. Data on diabetic kidney disease has been contributed by the SUMMIT Consortium and has been downloaded from mccarthy.well.ox.ac.uk

(http://mccarthy.well.ox.ac.uk/publications/2018/SUMMIT_DKD/) on 9/02/2022.

Author Contributions

A.H performed the analysis and wrote the manuscript. M.L.M. conceived the hypothesis, prepared data, contributed to discussion and reviewed/edited manuscript. C.W helped conceive the analysis pipeline, supervised the analysis, contributed to discussion and reviewed/edited manuscript. P.B, F.J.C, S.T.C, J.J.C, R.N.D, D.D, E.A.D, J.E.D, K.C.D, T.W.J, F.H.M, S.M.M, A.N contributed to data collection and critical review of the manuscript. W-M.C, S.O-G, S.S.R contributed to genetic analyses.

Conflict of Interest statement

CW and AH are partially supported by GSK and CW is partially supported by MSD. These funders had no involvement in this study. CW also holds a part time position with GSK.

Funding

AH is funded by the Engineering and Physical Sciences Research Council (EPSRC; EP/R511870/1) and the Medical Research Council (MRC; MC UU 00002/4). CW is funded by the Wellcome Trust (WT220788), the Medical Research Council (MRC; MC UU 00002/4), GSK and MSD and supported by the NIHR Cambridge BRC (BRC-1215-20014). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research was funded in

whole, or in part, by the Wellcome Trust (WT220788). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The AddIT study was funded by Diabetes UK, Juvenile Diabetes Research Foundation, the British Heart Foundation and in Canada the JDRF-Canadian Clinical Trial Network (CCTN), the Canadian Diabetes Association and the Heart and Stroke Foundation Canada. The Oxford Regional Prospective Study was funded by Diabetes UK. The NFS is funded by the Juvenile Diabetes Research Foundation, the Wellcome Trust and Diabetes UK. We acknowledge the study field workers, paediatricians, physicians, and diabetes nurse specialists involved in the ORPS/NFS/AddIT studies and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. The funding for the genotyping and Q/C for the JDRF Diabetic Nephropathy Collaborative Research Initiative was through JDRF grant #17-2013-7.

References

1. Alicic RZ, Rooney MT, Tuttle KR. Diabetic Kidney Disease: Challenges, Progress, and Possibilities. *Clin J Am Soc Nephrol CJASN*. 2017 Dec 7;12(12):2032–45.
2. Bjerg L, Hulman A, Carstensen B, Charles M, Witte DR, Jørgensen ME. Effect of duration and burden of microvascular complications on mortality rate in type 1 diabetes: an observational clinical cohort study. *Diabetologia*. 2019 Apr;62(4):633–43.
3. Matsushita K, Coresh J, Sang Y, Chalmers J, Fox C, Guallar E, et al. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data. *Lancet Diabetes Endocrinol*. 2015 Jul;3(7):514–25.
4. Marcovecchio ML, Chiesa ST, Armitage J, Daneman D, Donaghue KC, Jones TW, et al. Renal and Cardiovascular Risk According to Tertiles of Urinary Albumin-to-Creatinine Ratio: The Adolescent Type 1 Diabetes Cardio-Renal Intervention Trial (AddIT). *Diabetes Care*. 2018 Sep;41(9):1963–9.
5. Benitez-Aguirre PZ, Marcovecchio ML, Chiesa ST, Craig ME, Wong TY, Davis EA, et al. Urinary albumin/creatinine ratio tertiles predict risk of diabetic retinopathy progression: a natural history study from the Adolescent Cardio-Renal Intervention Trial (AddIT) observational cohort. *Diabetologia*. 2022 May 1;65(5):872–8.
6. Cole JB, Florez JC. Genetics of diabetes mellitus and diabetes complications. *Nat Rev Nephrol*. 2020 Jul;16(7):377–90.
7. Amin R, Widmer B, Prevost AT, Schwarze P, Cooper J, Edge J, et al. Risk of microalbuminuria and progression to macroalbuminuria in a cohort with childhood onset type 1 diabetes: prospective observational study. *BMJ*. 2008 Mar 29;336(7646):697–701.
8. Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB, et al.

- Genome-Wide Association Scan for Diabetic Nephropathy Susceptibility Genes in Type 1 Diabetes. *Diabetes*. 2009 Feb 27;58(6):1403–10.
9. Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, Isakova T, et al. New Susceptibility Loci Associated with Kidney Disease in Type 1 Diabetes. *PLOS Genet*. 2012 Sep 20;8(9):e1002921.
 10. Sandholm N, McKnight AJ, Salem RM, Brennan EP, Forsblom C, Harjutsalo V, et al. Chromosome 2q31.1 Associates with ESRD in Women with Type 1 Diabetes. *J Am Soc Nephrol*. 2013 Oct 1;24(10):1537–43.
 11. Salem RM, Todd JN, Sandholm N, Cole JB, Chen WM, Andrews D, et al. Genome-Wide Association Study of Diabetic Kidney Disease Highlights Biology Involved in Glomerular Basement Membrane Collagen. *J Am Soc Nephrol*. 2019 Oct 1;30(10):2000–16.
 12. Teumer A, Tin A, Sorice R, Gorski M, Yeo NC, Chu AY, et al. Genome-wide Association Studies Identify Genetic Loci Associated With Albuminuria in Diabetes. *Diabetes*. 2015 Dec 2;65(3):803–17.
 13. Böger CA, Chen MH, Tin A, Olden M, Köttgen A, Boer IH de, et al. CUBN Is a Gene Locus for Albuminuria. *J Am Soc Nephrol*. 2011 Mar 1;22(3):555–70.
 14. Sandholm N, Forsblom C, Mäkinen VP, McKnight AJ, Österholm AM, He B, et al. Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes. *Diabetologia*. 2014 Jun 1;57(6):1143–53.
 15. Marcovecchio ML, Dalton RN, Daneman D, Deanfield J, Jones TW, Neil HAW, et al. A new strategy for vascular complications in young people with type 1 diabetes mellitus. *Nat Rev Endocrinol*. 2019 Jul;15(7):429–35.
 16. Schultz CJ, Neil HA, Dalton RN, Dunger DB, Oxford Regional Prospective Study Group. Risk of nephropathy can be detected before the onset of microalbuminuria during the early years after diagnosis of type 1 diabetes. *Diabetes Care*. 2000 Dec;23(12):1811–5.
 17. Marcovecchio ML, Dalton RN, Prevost AT, Acerini CL, Barrett TG, Cooper JD, et al. Prevalence of abnormal lipid profiles and the relationship with the development of microalbuminuria in adolescents with type 1 diabetes. *Diabetes Care*. 2009 Apr;32(4):658–63.
 18. Marcovecchio ML, Woodside J, Jones T, Daneman D, Neil A, Prevost T, et al. Adolescent Type 1 Diabetes Cardio-Renal Intervention Trial (AdDIT): urinary screening and baseline biochemical and cardiovascular assessments. *Diabetes Care*. 2014;37(3):805–13.
 19. Marcovecchio ML, Chiesa ST, Bond S, Daneman D, Dawson S, Donaghue KC, et al. ACE Inhibitors and Statins in Adolescents with Type 1 Diabetes. *N Engl J Med*. 2017 Nov 2;377(18):1733–45.
 20. Dunger DB, Schwarze CP, Cooper JD, Widmer B, Neil H a. W, Shield J, et al. Can we identify adolescents at high risk for nephropathy before the development of microalbuminuria? *Diabet Med J Br Diabet Assoc*. 2007 Feb;24(2):131–6.
 21. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016 Oct;48(10):1284–7.
 22. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016 Oct;48(10):1279–83.
 23. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009 Jun 29;338:b2393.
 24. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011 Dec 12;45:1–67.
 25. Rubin DB. Multiple Imputation for Nonresponse in Surveys | Wiley [Internet]. Wiley.com. 1987 [cited 2022 Jan 19]. Available from: <https://www.wiley.com/en-gb/Multiple+Imputation+for+Nonresponse+in+Surveys-p-9780471655749>
 26. Bürkner PC. brms: An R Package for Bayesian Multilevel Models Using Stan. *J Stat Softw*. 2017 Aug 29;80:1–28.
 27. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association

- studies. *Nat Genet.* 2012 Jun 17;44(7):821–4.
28. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet.* 2012 Dec;44(12):1294–301.
 29. Hutchinson A, Watson H, Wallace C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Comput Biol.* 2020 Apr 13;16(4):e1007829.
 30. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Ser B Stat Methodol* [Internet]. [cited 2020 Jul 28];n/a(n/a). Available from: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12388>
 31. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the “Sum of Single Effects” model [Internet]. 2021 Nov [cited 2022 Jan 19] p. 2021.11.03.467167. Available from: <https://www.biorxiv.org/content/10.1101/2021.11.03.467167v1>
 32. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015 Oct;526(7571):68–74.
 33. Sandholm N, Van Zuydam N, Ahlqvist E, Juliusdottir T, Deshmukh HA, Rayner NW, et al. The Genetic Landscape of Renal Complications in Type 1 Diabetes. *J Am Soc Nephrol JASN.* 2017 Feb;28(2):557–74.
 34. Zuydam NR van, Ahlqvist E, Sandholm N, Deshmukh H, Rayner NW, Abdalla M, et al. A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects With Type 2 Diabetes. *Diabetes.* 2018 Jul;67(7):1414.
 35. ALFA: Allele Frequency Aggregator [Internet]. [cited 2022 Jan 20]. Available from: <https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>
 36. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics.* 2015 Nov 1;31(21):3555–7.
 37. Pazoki R, Evangelou E, Mosen-Ansorena D, Pinto RC, Karaman I, Blakeley P, et al. GWAS for urinary sodium and potassium excretion highlights pathways shared with cardiovascular traits. *Nat Commun.* 2019 Aug 13;10(1):3653.
 38. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013 Jun;45(6):580–5.
 39. Wu NN, Zhao D, Ma W, Lang JN, Liu SM, Fu Y, et al. A genome-wide association study of gestational diabetes mellitus in Chinese women. *J Matern-Fetal Neonatal Med Off J Eur Assoc Perinat Med Fed Asia Ocean Perinat Soc Int Soc Perinat Obstet.* 2021 May;34(10):1557–64.
 40. Schlicht K, Nyczka P, Caliebe A, Freitag-Wolf S, Claringbould A, Franke L, et al. The metabolic network coherence of human transcriptomes is associated with genetic variation at the cadherin 18 locus. *Hum Genet.* 2019;138(4):375.
 41. Thorn LM, Forsblom C, Fagerudd J, Thomas MC, Pettersson-Fernholm K, Saraheimo M, et al. Metabolic syndrome in type 1 diabetes: association with diabetic nephropathy and glycemic control (the FinnDiane study). *Diabetes Care.* 2005 Aug;28(8):2019–24.
 42. Palaniappan L, Carnethon M, Fortmann SP. Association between microalbuminuria and the metabolic syndrome: NHANES III*: *Am J Hypertens.* 2003 Nov 1;16(11):952–8.
 43. Kato K, Hansen L, Clausen H. Polypeptide N-acetylgalactosaminyltransferase-Associated Phenotypes in Mammals. *Mol Basel Switz.* 2021 Sep 10;26(18):5504.
 44. Peng C, Togayachi A, Kwon YD, Xie C, Wu G, Zou X, et al. Identification of a novel human UDP-GalNAc transferase with unique catalytic activity and expression profile. *Biochem Biophys Res Commun.* 2010 Nov 26;402(4):680–6.
 45. Memarian E, 't Hart LM, Slieker RC, Lemmers RFL, van der Heijden AA, Rutters F, et al. Plasma protein N-glycosylation is associated with cardiovascular disease, nephropathy, and retinopathy in type 2 diabetes. *BMJ Open Diabetes Res Care.* 2021 Oct 13;9(1):e002345.

Tables

- **Table 1: Clinical characteristics for the subjects included in our analysis.** Values are expressed as mean±SD for normally distributed continuous variables, median (IQR) for non-normally distributed variables and percentage for categorical variables.
- **Table 2: SNPs with suggestive evidence of association ($p < 1 \times 10^{-6}$) in any of the three GWAS phenotypes.** For each index SNP, the table contains the rsID, the chromosome, the base position (bp, hg19), the alleles (reference > effect), the estimated effect size (beta), the standard error of the estimated effect size (SE), the observed and adjusted p-values and the closest protein coding gene. Note that rs145715205 appears twice, once for the trajectory and once for the latest phenotype.
- **Table 3: Replication results for rs150766792 from the Type 1 Diabetes Knowledge Portal (T1DKP; <https://t1d.hugeamp.org/>) in studies where subjects were of European ancestry.** The effect allele for rs150766792 was G and the reference allele was A.

Figure legends

- **Fig. 1: Figure to illustrate modelling ACR trajectories to derive the trajectory phenotype.** In plots of covariate-adjusted ACR measurements against duration of diabetes, we show the observed trajectory (line between data points for each individual), the modelled trajectory without borrowing information across individuals (straight “best fit” line fitted to data points for each individual) and the modelled trajectory with borrowing across individuals (from which the slope equals the “trajectory phenotype”). Panels (A) and (B) highlight trajectories for individuals with many data points available, where the borrowing only attenuates the slope of the trajectory slightly. Panels (C) and (D) highlight trajectories for individuals with only

three data points available, where the borrowing attenuates the slope of the trajectory more heavily. The grey lines represent a random sample of observed trajectories. The average trajectory across all subjects is shown in black. The number of visits and the attenuation between the trajectory modelled with borrowing and without borrowing is given in each panel. Attenuation was calculated as $1 - \text{slope of trajectory with borrowing} / \text{slope of trajectory without borrowing}$.

- **Fig. 2: Figure to illustrate observed trajectories for individuals with the highest and lowest phenotype values.** In plots of covariate-adjusted ACR measurements against duration of diabetes, we highlight the observed trajectories for individuals with the highest (A) and lowest (B) phenotype value for each of the three phenotypes. The grey lines show a random sample of observed trajectories.
- **Fig. 3: Manhattan plots of GWAS results for (A) latest ACR phenotype (B) average ACR phenotype and (C) trajectory phenotype.** Solid line on Manhattan plots show genome-wide significance threshold ($p = 5 \times 10^{-8}$) and dashed line shows suggestive significance threshold ($p = 1 \times 10^{-6}$). The y-axis is the $-\log_{10}$ transformed p-value and the x-axis is the genomic position.
- **Fig. 4: Manhattan plot for the *GALNTL6* gene region (chr54:167000000-176000000).** The index SNP in the region is highlighted and labelled in red. The two other SNPs included in the 95% credible set are highlighted and labelled in blue. Genes and genomic position (hg19) along chromosome 4 are shown in the bottom panel.
- **Fig. 5: Manhattan plot for the *CDH18* gene region (chr5:16000000-25000000).** The index SNP in the region is highlighted and labelled in red. Genes and genomic position along chromosome 5 are shown in the bottom panel (hg19).