

IMI1 Final Project Report Public Summary

Project Acronym: DIRECT

Project Title: Diabetes REsearCh on patient
sTratification

Grant Agreement: 115317

Project Duration: 01/02/2012 – 31/07/2019

1. Executive summary

The IMI-DIRECT project was 80 months in duration and aimed to investigate heterogeneity in two key areas: glycaemic deterioration prior to and after the onset of diabetes, and glycaemic response to diabetes treatments, including drug treatment and obesity surgery. Over the course of the project the consortium generated in excess of 40TB of clinical, imaging and molecular data on 3000 participants at baseline and with two follow up visits over 3-4 years. To investigate therapeutic response a prospective study of patients starting GLP-1RA was undertaken, with physiological studies undertaken to investigate beta-cell function and glucose sensitivity. In addition, large observational and trial datasets, linked to genomics, were explored and genetic biomarkers for metformin intolerance, sulphonylurea and GLP-1RA response were identified which may be beneficial for genotype guided prescribing. With ongoing funding from EFPIA the consortium will continue in a legacy phase for at least two years, and data made available by request to other research groups. A number of exciting outputs have been generated to date, but with the depth of data generated and the longitudinal nature of these measures, the resource will become an increasing asset for both the IMI-DIRECT consortium and the wider research community.

1.1 Project rationale and overall objectives of the project

The overarching aims of the IMI-DIRECT consortium are to identify biomarkers that address current bottlenecks in diabetes drug development and to develop a stratified medicines approach to treatment of 'type 2 diabetes' (T2D) with either existing or novel therapies. T2D is a highly heterogeneous condition largely because when hyperglycaemia is discovered it is a diagnosis made after exclusion of other known causes. The phenotype of people who develop diabetes is highly variable, as is the rate at which their subsequent diabetes progresses, how they respond to diabetes therapy and who develops micro- and macrovascular complications. This variable development, progression and treatment response of diabetes is likely to reflect variation in underlying disease processes that drive the development of diabetes and its subsequent progression. The DIRECT consortium addresses two areas of heterogeneity: the rate at which people's glycaemia deteriorates either from prediabetes to diabetes, or through diabetes; and the extent to which certain therapeutic interventions result in improvement in glycaemia. The overall objectives are to 1) identify biomarkers that can stratify populations by risk of progression or most likely to respond to particular therapies to allow a precision medicine approach to diabetes management and diabetes risk reduction; and 2) to develop new mechanistic hypotheses for disease progression or response to therapeutic intervention leading to new therapies or to a better use of existing therapies. The ultimate aim of DIRECT is to progress the field of patient stratification with biomarkers arising from IMI-DIRECT being taken forward to one or more prospective clinical trials as a basis for approval by authorities and subsequent clinical use.

1.2 Overall deliverables of the project

In this 90-month programme of work, there were 7 active work packages. The overall deliverables were:

1. The prospective collection of large well-phenotyped cohorts of glycaemic deterioration with 4 years follow up, drug response and response to obesity surgery;

DIRECT

2. The assimilation of pre-existing data and samples available to consortium members;
3. Extensive physiological and genomic analysis of all available samples, including longitudinal samples;
4. Hosting of all data in a single secure server to enable high performance analytics;
5. Building a central biobank to enable future biomarker discovery and replication;
6. Integrative analysis across phenotype and multiomic levels to identify biomarkers of drug response and glycaemic deterioration;
7. Public and patient engagement;

1.3 Summary of progress versus plan since last period

Following the second review by IMI with independent experts in September 2017, it was agreed that resources and effort for this last remaining period should focus on maximising the potential for biomarker discovery. As a result, funds that had been held back for a potential clinical trial were re-allocated to extend our biomarker assays on 'follow-up' samples and to support analysis of the generated data. In addition, we have focused on developing plans for a Legacy Consortium, to continue after DIRECT for at least 2 years in the first instance. This has required legal agreements and arrangements to fund the ongoing maintenance of the database and biobank.

WP1. In the reporting period, two further amendments (amendment no.9 and no. 10) were submitted to IMI and approved. Amendment no. 9 requested the extension of the DIRECT project from 84 to 90 months with a new project end date of 31st July 2019. Amendment no. 10 implemented a final comprehensive re-allocation of academic funds affecting 11 beneficiaries. Furthermore, EFPIA partners SAD, ELI LILLY and BI agreed again to convert in-kind resources into 1.3 mill Euro cash for financial support of the academic partner UNIGE for RNA sequencing and to enable intensive metabolomics analysis of WP2 samples by a subcontractor (Metabolon).

WP2. All clinical data collection for WP 2.1 was completed by Q4 of 2018; all 2.1 samples were delivered to the respective central labs for analyses. All assays have been performed, including final baseline biomarker data (miRNA within WP2.2, the follow-up O-link proteomic data, and the intestinal microbiome sequence data) and fsOGTT modelling is complete. The clinical database has been locked at the end of project funding (31/7/19). A plan is in place to collect participant data for 6 years beyond project end.

WP3. We have completed a candidate gene, GWAS and rare variant analysis across IMI-DIRECT cohorts, the AWARD studies (Dulaglutide, Lilly) and the HARMONY studies (Albiglutide, GSK). We have identified genome-wide significant variants in the GLP-1R and Arrestin-B1 that alter response to GLP-1RA. Functional work supports the mechanism. We have also completed a meta-GWAS of glycaemic response to sulphonylureas that includes the DIRECT generated data. Here we find a genome wide hit in SLCO1B1 – a drug transporter well known to transport Statins, but only recently reported to transport sulphonylureas. For remission of diabetes following obesity surgery we have developed a multiomic

model that includes a variety of measured proteins and metabolites that considerably outperforms the predictive value of the current clinical model.

WP4. Following the reallocation of funds to carry out additional omics there has been considerable activity during this reporting period. We have undertaken RNA-seq, Metabolomics (Biocrates and Metabolon) on all follow up samples (18 and 36/48 months), O-link proteomics on the 36/48 month time points and deep sequencing of the gut microbiome on the 36/48 month stool samples. Analysis is complete although QC is still ongoing. In addition to analysis of the DIRECT prospective cohorts, we have also undertaken GWAS of the HARMONY-Outcomes study (albiglutide, WP3) and the LOOKAHEAD study (Diabetes progression, WP2).

WP5. In preparation for the Legacy phase of DIRECT, a major normalisation and harmonisation effort of the non-eCRF data has been initiated in order to allow for the design, implementation and optimisation of a solution to store all data in a centralised and easy-to-query database. For this purpose, all data type files have been converted to a uniform format (csv) and all variables have been extracted. Additionally, all metadata and “README” files have been collected to provide analysts and data managers with the appropriate description of variables and, when needed, experimental information relative to data generation. As the number of DIRECT participants involved in collecting and quality assessing data has been very large over the long period this has been a major effort. Another important element in the Legacy preparation has been to make sure that computational infrastructure and resources are maintained in the post-DIRECT timeframe to allow for the continued operation of the analysis platform for DIRECT researchers to perform the complex analysis of DIRECT data. Finally, the DIRECT data has been CDISC aligned in line with the RHAPSODY project to enable data sharing between IMI projects.

WP6. The initial aim of WP6 was to develop commercial assay prototypes for measuring metabolomic and proteomic biomarkers discovered in DIRECT and apply these tests in the clinical study. Indeed, two of the three remaining milestones & deliverables for the last reporting period addressed the establishment of commercial assay prototypes for newly discovered protein and metabolite biomarkers. However, due to re-prioritisation in the project, the search for new biomarkers is still an ongoing process.

Subsequent to the decision to de-prioritise a clinical trial within DIRECT, WP6 instead focused on applying its technology portfolio to generate more insights from samples collected at 36- and 48-month follow-up. Within this effort, WP6 applied assays and technologies (e.g. ProteinSimple) which were evaluated as suitable platforms for commercial assay prototypes for measuring DIRECT biomarkers.

WP9. With the allocation of additional funds, WP9 has been working with other DIRECT partners to draft and negotiate the Data Legacy Contract to enable the continued use of the DIRECT data once the project funding comes to an end. There has been considerable discussion and legal analysis as to how the GDPR applies and how this affects the way the contracts should be drafted. WP9 has also compiled the final patient yearly newsletter for the project. Translations and dissemination of the newsletter are undertaken by the respective study teams in each country. In addition, two publications are in

preparation, which will contribute to the four that have already been published on the governance issues associated with the DIRECT project.

WP2,4,5,6. All data analysis has been undertaken collaboratively across work packages. In the last 18 months, analysts have had weekly calls, joined by Principal Investigators once a month. Analysis-specific working groups have been formed and these have reported into the monthly calls. Analyses are coordinated to minimize overlap and ensure consistency of methods and outputs by ensuring analysts interested in overlapping research questions work together. As a result, the outputs (in terms of publications and presentations at scientific meetings) have increased, although we anticipate that most major outputs will occur during the 2 years following the official end of the IMI-DIRECT funding.

1.4 Significant achievements since last report

WP1. The sustainability plan of the DIRECT consortium has been further developed in PY7. After EFPIA partners SAD, ELI LILLY, NOVO and BI agreed to provide funds for the continuation of the consortium for the next two years after project end on 31st July 2019, the preparation of the DIRECT Legacy Agreement and the DIRECT Sponsorship Agreement under the leadership of UOXF and SAD, respectively, commenced. All DIRECT participants except SERVIER indicated they wished to remain a partner of the new DIRECT Legacy Consortium by signing the DIRECT Legacy Agreement. The decision-making body of the DIRECT Legacy Consortium shall be the DIRECT Legacy Board composed of one representative from each participant. The agreements have been analysed to ensure that they are compliant with the GDPR.

WP2.

Data: The collection of all clinical data for WP 2.1 was completed by Q4 of 2018, with all 2.1 samples being delivered to the respective central labs. There was a final data freeze on 31st July 2019. Please also refer to D2.1.4 and D2.2.4 for detail. Dietary Food Frequency Questionnaires (FFQ) analysis on baseline, 18m and most of the 36m data is complete. The outstanding data (incl. 48m FFQ data) will be uploaded by Q4 2019.

Progression models: Final models were agreed upon for defining the progression of diabetes, including OGTT modelling for all time-points within DIRECT - please also refer to D2.3.3 and D2.5.2 for details. In addition, progression modelling within the LookAhead study - incorporating their GWAS data with other diabetes progression GWAS data available in DIRECT (Go-DARTS, DCS and the prospective WP2.2 cohort) - is ongoing.

Data Legacy: The possibility of collecting key data from WP2 study participants (WP2.1 and 2.2) on a regular basis beyond the formal end of the DIRECT study will greatly enrich the present DIRECT cohort data. The development of established data collection systems and protocols needed to enable clinical data retrieval for all consented WP2 participants at most participating WP2 sites, from study-end (31st July 2019) through to December 2026, has now been completed. This includes the further development and maintenance of a DIRECT legacy database at DTU, which is to be updated biennially with data retrieved from each participating centre. Please also refer to D2.7.1 for details.

WP3. Analyses in this period have focused on the GLP-1RA genetic data. We have identified genetic variants in the GLP-1RA and Arrestin- β 1 that alter response to GLP-1RA. The clinical effect is modest. We have supporting functional data that establishes a novel role for the arrestin-b1 variant in altering GLP-1R internalisation.

WP4. In this reporting period, WP4 has successfully coordinated the large-scale analysis of the WP2 cohort follow-up samples following the approved budget reallocation at the start of 2019. This includes RNA-seq on all follow up time points, metabolomics on all time points (Biocrates P150; metabolon); O-link proteomics and intestinal microbiome deep sequencing on the 36/48 month samples. In addition, the baseline microbiome sample sequencing and QC has been completed and is currently being analysed.

WP5. In this reporting period WP5 has successfully made significant achievements regarding the repository and the legacy status of the DIRECT database. We have re-organized the repository directories and collected metadata from data submitters. Furthermore, we have executed a comprehensive clean-up of redundant, out-of-date, and wrong data in collaboration with data submitters or data contacts. The “deleted/removed” data has been placed in “NOT_FOR_LEGACY” folders in the sub-directories of the repository. Finally, the full repository has been cloned into /home/Data/Repository/tmp/Rep with all delimited files, converted and cleaned into parsable csv files.

WP6. Proteomics activities focused on conducting and completing the immunoassay analysis of proteins in plasma samples from WP2. By using several immunoassay platforms such as Olink, ProteinSimple and Luminex, we have analysed 3000 baseline samples on nearly 500 protein biomarkers selected by DIRECT, as well as 2500 longitudinal samples on about 200 of these proteins. In addition to this, we further analysed 400 proteins in 400 plasma samples from WP3.5 using Olink assays. The first insights and indications from these studies were presented at the 78th ADA conference and at 2 international proteomics meetings.

Proteomics data were also used to conduct analysis of pre-analytical variables. This investigated the extent to which the parameters concerning the samples had an impact on the measured protein levels. This survey brought about important information on variables such as clinical centre, age of sample and the time point of the sample collection that will now assist the analysis of the data in relation to clinical parameters. The pre-analytical variables affected the data across different proteomics platforms and will allow a more informed analysis of the data.

We also focused on the data cleaning and normalization of the existing targeted and non-targeted metabolomics data sets and in collaboration with WP4, HMGU lead the study design and measurements of the WP2 follow-up samples. These measurements are ongoing and should be finished in October 2019.

WP9. Two papers by Shah *et al*, 2019 (1, 2) were published on our patient survey data that asked patients on their preferences about who could access their data, and which has formed the basis for designing the Legacy Database governance. The participant survey had 855 responses in total from DIRECT data collection centres with a high response rate of 80% returned and completed. The analysis of the results shows that overall 97% of participants were happy to share their DIRECT consortium data beyond the project. The most important points for participants in how data access is managed were that the database must be secure - 80% rated this as highly important, and 69% stated that DIRECT researchers should

monitor use as highly important. This was an important piece of research before the development of the Data Legacy contract.

WP9 have also assisted in the drafting and negotiation of the Data Legacy Contract to enable the use of the DIRECT data once the project funding comes to an end. There has resulted in considerable discussion and legal analysis as to how the GDPR applies and how this affects the way the contracts should be drafted. Based on this, we have two publications in preparation, which adds to the other three papers that have already been published. WP9 has also compiled the final patient yearly newsletter for the project, which describes individual rights under the GDPR, key achievements of the project and what will happen with the data collected and generated in the DIRECT project.

WP2,4,5,6. In the combined cross-work package analyses of the prospective cohorts, we have achieved a number of successes that are largely summarized in section 1.5.

1.5 Scientific and technical results/foregrounds of the project

1.5.1 Overview. The IMI-DIRECT cohorts are the most comprehensively phenotyped cohorts to investigate diabetes and metabolic disease to date. The IMI-DIRECT consortium has generated 40 TB of data, consisting of CRF, imaging and phenotype data, multiomics including stool sequencing and RNA-seq and has over 300,000 samples stored in its biobank. A web accessible interface has been implemented to allow for recruitment centres to upload and manage clinical and eCRF data for all recruited patients. Access to the interface is exclusively granted to DIRECT approved personnel and is the only point of entry for the data server. We have a legacy plan in place to ensure ongoing use by the consortium and the wider scientific community. As the prospective studies only completed in December 2018, much of the multiomic data have only recently been generated. We anticipate our major scientific results will be generated in the next 2 years; however, we have generated exciting results to date, which are summarised here.

1.5.2 eCRF Figure 1 (below) summarises the clinical and eCRF data collected in DIRECT's data warehouse during the duration of the project as of March 2019. The interface and its statistical module has also played an important role in project management where all partners have been able to track the recruitment and the corresponding growth in the data produced from the samples collected in the clinics. Statistics have also been produced for each study centre over the period (not shown here).



Fig 1: eCRF data in the DIRECT database (June 2019)

This interface will remain operative during the Legacy phase as it will be used to collect and store data from the agreed additional visits scheduled. As the precise length of the additional period will depend on additional funding, we aim for having the interface operational even after the currently agreed legacy phase to further increase the value of the DIRECT data.

1.5.3. Proteomics. The aim of the DIRECT WP6 proteomics activities was to discover protein biomarkers and to develop assay prototypes for a clinical trial. In order to build a framework for the transition of discovery to clinical assays, KTH has worked on the means to validate the data and build dual binder assays for a quantitative measurement of protein biomarkers.

KTH has published a technical research article on validating the antibody selectivity when using these reagents to capture a protein from blood plasma (3). The paper describes the use of mass spectrometry for the analysis of antibodies binding to protein biomarker candidates found in DIRECT studies around WP3.4 and WP3.5. Here, we demonstrate the on-target detection of proteins such as IGFBP2 and CCL16 and discover that interactions between proteins (e.g. IGFBP2 and IGF1) can be detected in plasma. The validation of selectivity is a prerequisite to develop precision assays for clinical translation and utility.

KTH has further published a second technical proteomics research article that describes the workflow for developing dual binder assays for the analysis of plasma protein biomarkers, such as IGFBP2 and CCL16 (4). This study is the result of establishing a routine pipeline to select suitable pairs of antibodies in order to quantify a protein found in blood plasma by dual binder assays. We also compared the established assays with other technologies such as Olink and targeted mass spectrometry. In addition, we examined the longitudinal stability of protein levels in blood using non-DIRECT samples.

1.5.4 Results of ongoing joint analyses between WP2,3,4,5,6

1.5.4.1 Biomarkers of characteristics at recruitment

1.5.4.1.1 Machine learning based tool for the diagnosis of non-alcoholic fatty liver disease

Non-alcoholic fatty liver disease (NAFLD) is highly prevalent and causes serious health complications in type 2 diabetes (T2D) and beyond. Early diagnosis of NAFLD is important, as this can help prevent irreversible damage to the liver. Utilizing data from the IMI DIRECT consortium we developed a diagnostic tool for NAFLD by applying machine-learning approaches. Multiomics (genetic, transcriptomic, proteomic and metabolomic) and clinical data, including plasma biomarkers, measures of beta-cell function, insulin sensitivity, anthropometry, and lifestyle, comprised the key input variables. The outcome variable was MRI image-derived liver fat content, categorized into fatty ($\geq 5\%$) or normal ($< 5\%$) ($n=1514$ participants). Feature selection of different layers of omics data was performed using LASSO (least absolute shrinkage and selection operator) and Random Forest analysis was used to develop the algorithms. Several prediction models for NAFLD were developed using clinical and omics data separately or in combination. The 'Advanced' model (all omics and clinical variables) yielded a cross-validated area under the receiver operator characteristic curve (ROCAUC) of 84.3% (95% CI= 82.1%, 86%) - see Figure 2 below. For the 'pragmatic' model (only 9 clinically-accessible variables), the ROCAUC was 81.7% (95% CI= 80.6%, 82.6%).
Manuscript under consortium review.

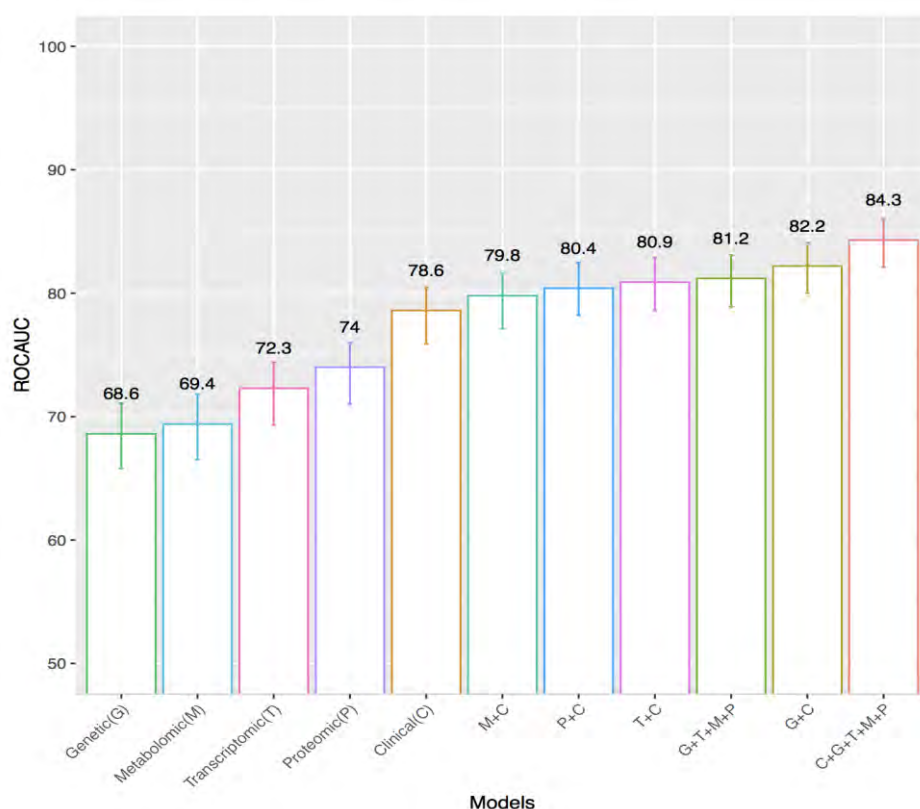


Fig 2: The area under the receiver operator characteristic curve (ROCAUC) results (y-axis) for the omics models separately (Genetic (G), Transcriptomic (T), Proteomic (P) and Metabolomic (M)) or combined with the Clinical (C) model (C+G, C+T, C+P and C+M) in the test data of IMI DIRECT

1.5.4.1.2 The role of physical activity in metabolic homeostasis before and after the onset of type 2 diabetes

It is well established that physical activity (PA), abdominal ectopic fat and glycaemic regulation are related, but the underlying structure of these relationships is unclear. The previously proposed twin-cycle (TC) hypothesis provides a mechanistic basis for impairment in glycaemic control through the interactions of substrate availability, substrate metabolism and abdominal ectopic fat accumulation. Here we hypothesise that the effect of PA in glucose regulation is mediated by the TC. We defined a structural equation model that describes the TC hypothesis and fitted this within the IMI-DIRECT dataset. A second model (TC-PA), to assess the extent to which the effects of PA in glycaemic regulation are mediated by the TC, was also fitted. Beta-cell function, insulin sensitivity and glycaemic control were modelled from frequently sampled 75g oral glucose tolerance tests (fsOGTT) and mixed-meal tolerance tests (MMTT) in participants without and with diabetes, respectively. Abdominal fat distribution was assessed using MRI, and PA through wrist-worn triaxial accelerometry. Results are presented as standardized beta coefficients (β), standard errors (se) and P-values, respectively.

The TC and TC-PA models showed better fit than null models (TC: $\chi^2=242$, $P=0.004$ and $\chi^2=63$, $P=0.001$ in cohort 1 and 2 respectively; TC-PA: $\chi^2=180$, $P=0.041$ and $\chi^2=60$, $P=0.008$ in cohort 1 and 2 respectively). Insulin sensitivity mediated the inverse association between PA and fasting glucose (cohort 1: $\beta=-0.21$, $se=0.03$, $P<0.001$; cohort 2: $\beta=-0.15$, $se=0.04$, $P<0.001$) and 2-hr glucose (cohort 1: $\beta=-0.17$, $se=0.02$, $P<0.001$; cohort 2: $\beta=-0.14$, $se=0.04$, $P<0.001$). These analyses support the mechanisms proposed in the twin-cycle model with the exception of the pancreatic fat – glucose sensitivity axis (*see figure 3*). They highlight mechanistic pathways through which PA potentially affects metabolic homeostasis. *This manuscript is currently under review at Diabetologia.*

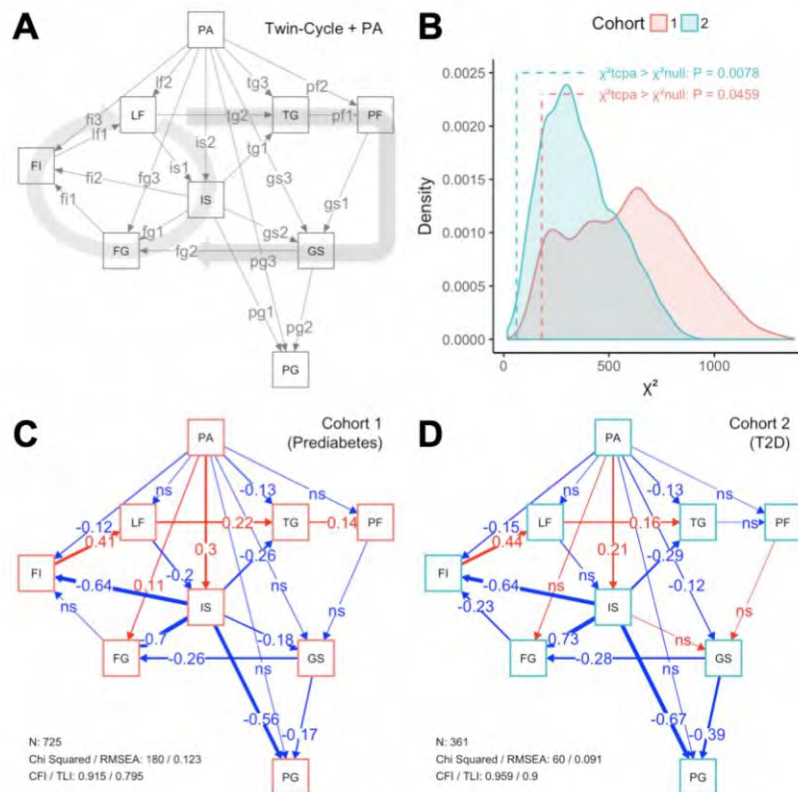


Fig 3: Twin-cycle structural equation model (TC) definition diagram, fit estimates, and effect estimate diagrams from a hypothesized model for the role of physical activity and liver fat in glycaemic control. Panel A: Model definitions; where manifest nodes are represented as squares, and arrows indicate regression coefficients pointing towards an outcome of a respective regression. Panel B: Model fit; density plot of model fit χ^2 from variable-randomized comparable structural equation models applied on respective dataset (10000 iterations). Solid vertical lines indicate TC model χ^2 , dashed lines indicate means χ^2 of all null iterations. Panel C and D: Effect estimate diagrams; of the defined model applied on cohort 1 (Prediabetes) and cohort 2 (T2D) respectively, where the arrow thickness is weighted by effect estimate magnitude, and colours red and blue indicate +ve and -ve estimates, respectively. All continuous variables are normally transformed and adjusted for: age, sex, metformin treatment (cohort 2), study centre, total energy-, carbohydrate-, fat-, and protein-intake. PA: Physical activity, FG: Fasting glucose, 2G: 2-hr glucose, IS: Oral glucose insulin sensitivity, LF: Liver fat, PF: Pancreatic fat, FI: Fasting insulin secretion rate, GS: Glucose sensitivity (insulin secretion per glucose).

1.5.4.1.3 Diet, the metabolome and metabolic health

Dietary advice remains the cornerstone of the prevention and management of type 2 diabetes. However, understanding the efficacy of dietary interventions is confounded by gross mis-reporting of current dietary assessment tools, especially in overweight individuals. Here, we investigate if plasma dietary metabolite profiling has utility in understanding glycaemic deterioration in people at risk or living with type 2 diabetes. We analysed baseline data and 18-month follow up in WP2.1 (healthy and pre-diabetic individuals) and WP2.2 (individuals with newly onset of type 2 diabetes). A plasma dietary metabolite profile model (Tpred) was constructed using 113 targeted plasma metabolites. We used this to explore the relationship between diet and glycaemic deterioration in people at risk or living with type 2 diabetes.

A higher T_{pred} score was associated with healthier dietary patterns (p trend=0.001). A higher T_{pred} score was also associated with lower body adiposity and lipid profiles (HDL-cholesterol and triglycerides) in both cohorts. In WP2.2 T_{pred} score was associated with a lower liver fat % ($\beta=-0.2$, $p<.0001$), and HbA1c ($\beta=-0.9$, $p=0.02$), fasting glucose ($\beta=-0.2$, $p=0.01$). Longitude data show a higher T_{pred} score also predicted favourable body compositions (BMI, waist circumference) in both cohorts at 18 months follow up and a lower fasting glucose ($\beta=-0.2$, $p=0.03$) and insulin ($\beta=-9.2$, $p=0.04$) in WP2.2. We conclude that plasma dietary metabolite profiling provides objective measures of diet intake, showing a relationship to glycaemic deterioration and cardio-metabolic health. *This manuscript is under consortium review; and will be submitted to Lancet Diabetes & Endocrinology.*

1.5.4.2 Physiological studies

1.5.4.2.1 Glycaemic deterioration in prediabetes

The trajectories of insulin sensitivity and beta-cell function parameters in WP2.1 are shown below in figure 4. Increase of insulin resistance, not associated with a significant increase in BMI, underlies the glucose tolerance deterioration. Beta-cell glucose sensitivity does not show deterioration, but insulin secretion at fixed glucose concentration and early secretion (rate sensitivity) do.

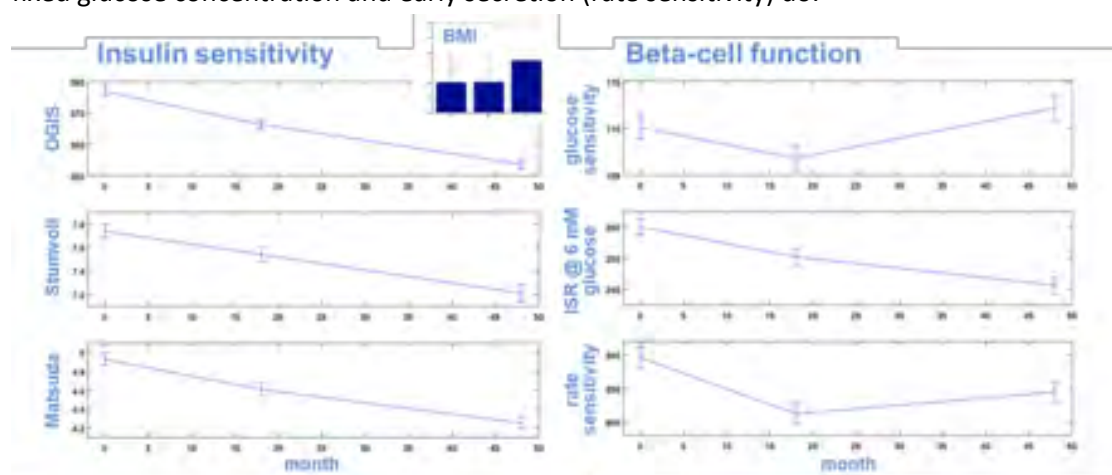


Fig 4: The trajectories of insulin sensitivity and beta-cell function parameters in WP2.1

The relationships between the mean OGTT glucose slope (by simple linear regression) and the changes from baseline to 48 months in insulin sensitivity (as the OGIS index) and two beta-cell function parameters are shown below. Figure 5 below shows that insulin sensitivity and insulin secretion at fixed glucose concentration are related to the mean OGTT glucose slopes. However, in the analogous plot using HbA1c slope, this relationship appears much weaker or absent.

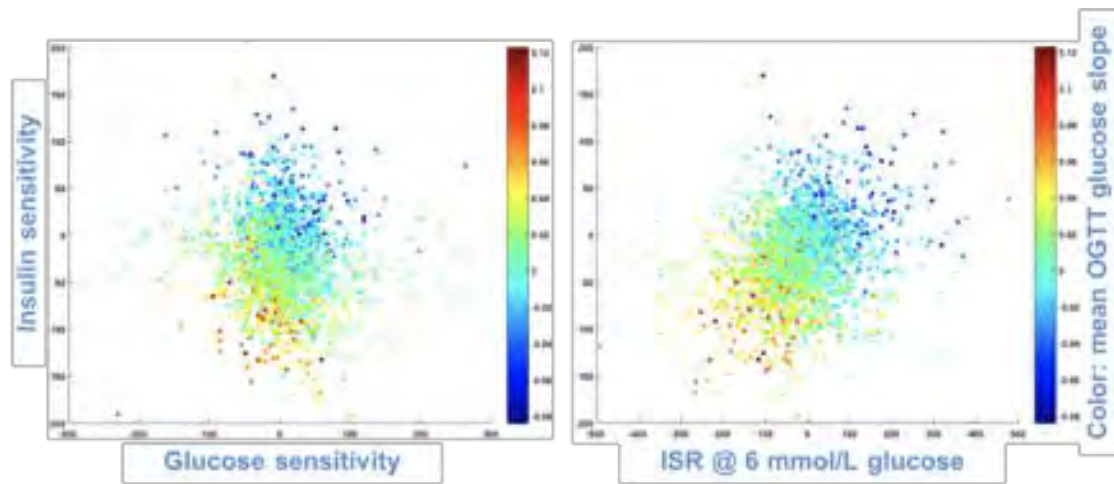


Fig 5: Relationships between the mean OGTT glucose slope and the changes from baseline to 48 months in insulin sensitivity (as the OGIS index) and glucose sensitivity and insulin secretion at 6 mmol/L glucose. The changes in insulin sensitivity and beta-cell function are shown on the axes, while the glucose slope is shown in colour. The highest slopes (red dots) are associated with decrease in insulin sensitivity (y-axis) and insulin secretion at fixed glucose (x-axis).

1.5.4.2.2 Glycaemic deterioration in patients with diabetes

We derived individual slopes for progression of HbA_{1c} and FPG in T2D, adjusted for BMI and antidiabetic treatment changes. As per figure 6 below, the analysis of HbA_{1c} slopes revealed that: 1) Greater HbA_{1c} progression is independently associated with baseline characteristics (fattier liver, and lower fasting glucose, age, GS and OGIS) as well as increments with time in triglycerides, ALT liver enzyme, β -cell glucose insensitivity and insulin resistance. 2) The ability to predict fast progression with these parameters was

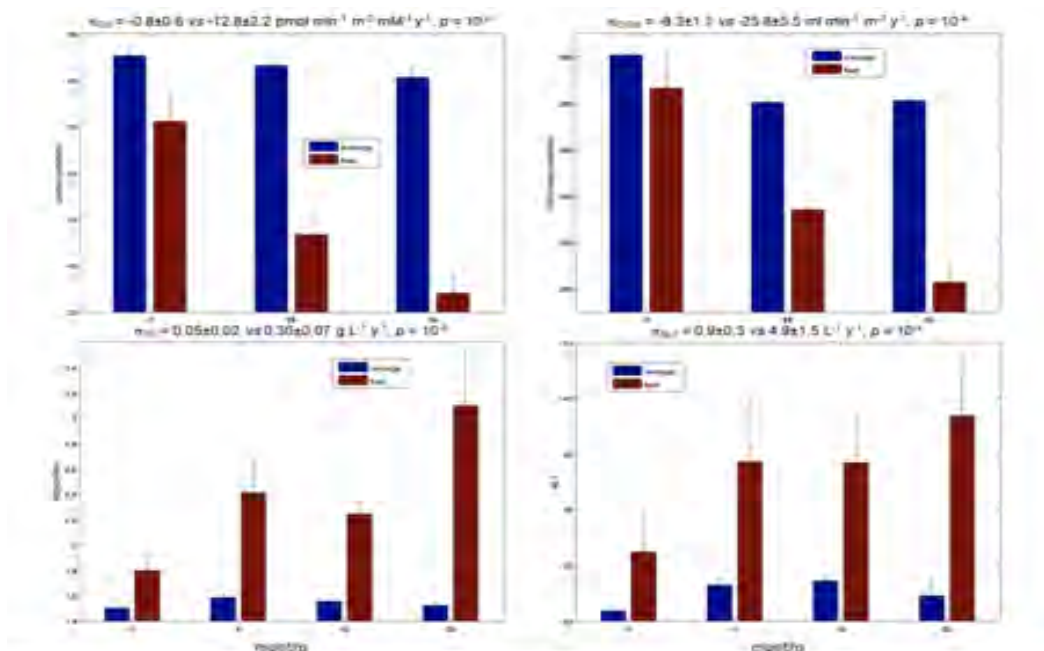


Fig 6: Summary of analyses of HbA_{1c} progression slopes in diabetics

high. 3) Fast vs average progression analysis was strongly consistent with the linear regression, as a much greater OGIS and GS deterioration was observed in fast progressors. 4) The strongest progression determinants are GS, OGIS and their longitudinal changes.

1.5.4.3 Single omics and glycaemic deterioration

1.5.4.3.1 Proteomics

Among 600 proteins the profiles of which were obtained for the baseline visit samples of 2233 prediabetic (WP2.1) and 793 diabetic patients (WP2.2) using various multiplexed assays, >300 proteins were found to be associated with diabetes related clinical traits. Some of those associations were replicated in the analysis of 2414 samples (1832 prediabetic and 582 diabetic) collected at the follow-up visit 36 or 48 months after the baseline visit. The confirmed proteins included LEP, PRSS8, IL-1ra, ACE2, IDUA, FGF-21, LPL, HAOX1 and RTN4R. Comparing the trends of the association between the two patient groups, we identified contrasting trend patterns for dozens of the pairs of proteins and the traits, which include steeper linear relationship between LEP and obesity related traits (e.g. waist, total abdominal adipose tissue (TAAT)) and milder between LEP and 2H OGIS in prediabetic people than diabetic patients.

Any proteins were pursued to find associated with the progression of diabetes after sample collection, which was evaluated by Roberto and his colleagues. No protein was discovered that had predicted the progression in WP2.1 (pre-diabetic patients). However, a few proteins (e.g. THBS2, TM, AMBP) were found to be associated with the progression in the diabetic patients (WP2.2).

1.5.4.4 Multiomic integration

1.5.4.4.1. Genomic architecture

The multiomics resource generated by the DIRECT project from blood and plasma samples from ~3,100 pre-diabetics and newly diagnosed patients with T2D was combined in an integrative framework to better explore the genetic basis of complex traits and T2D in particular. We previously reported on the exploration of QTL associations in each molecular data-type available, namely: transcriptomic (RNA-seq), proteomic (multiplexed immunoassays) and metabolomic.

In summary (*see table 1 below*), we identified 59,972 independent *cis*-eQTLs for 15,305 genes, 1,593 independent *cis*-pQTLs for 397 proteins. *Trans* genetics regulation of molecular phenotypes was also evaluated, identifying 2,320 *trans*-QTLs for 1,670 genes and 6,033 for 224 proteins (FDR 1%), in addition of the identification of 301 *metabo*-QTLs for 349 targeted and untargeted metabolites assays. Gene expression was the molecular phenotypes with the largest genetics regulation identify, with 97.07% of measured genes in DIRECT with a significant QTL influencing expression. This includes the mentioned *cis* and *trans* effects in gene expression phenotypes, *cis* effects in exon and splicing expression phenotypes, and *cis* effects on gene expression variance (*var*-eQTLs) and parent-of-origin effects (POE-eQTLs). In total, we identified 483 quantified genes (any phenotypes) without any type of eQTLs. In addition, we observed a substantial allelic heterogeneity, with 83.7% of genes and 84.1% of proteins associated with two or more variants.

Table 1: Number of QTLs identified for molecular phenotypes. We report the number of local (cis) genetics effects, as well as distal (trans) genetics effects in expression, and proteins. For metabolites without a known genomic location of references, we report all association. Secondary independent signals are also reported, identifying the number of phenotypes with more than one significant association, as well as the total number of QTL associations identified.

Phenotype	Num. phenotypes	1 st association	> 1 associations	Total num. of QTLs
Cis-eQTLs (genes)	16,205	15,305 (94.4%)	12,824	59,972
Cis-eQTLs (exons)	140,198 (15,712)	14,759 (93.9%)	12,477	59,952
Cis-seQTLs (splicing)	64,546	3,609 (5.59%)	298	3,978
Cis-pQTLs (targeted Proteins)	452	311 (68.80%)	156	614
Trans-eQTLs (genes)	15,114	1,670 (11.04%)	344	2,320
Trans-pQTLs (targeted proteins)	452	226 (49.77%)	158	6,153
metaboQTLs (targeted etabolites)	116	70 (52%)	33	103
metaboQTLs (untargeted metabolites)	233	103 (44%)	96	200

1.5.4.4.2. Clustering on baseline characteristics and progression of diabetes

We have undertaken non-linear dimensionality reduction using DDR tree on the baseline phenotypes available in WP2.2. We have then explored the molecular signature and association with diabetes progression for those individuals in each tree-end. This has enabled us to tease out aetiological processes for diabetes aetiology and how these associate with progression of diabetes (see figure 7 below for summary).

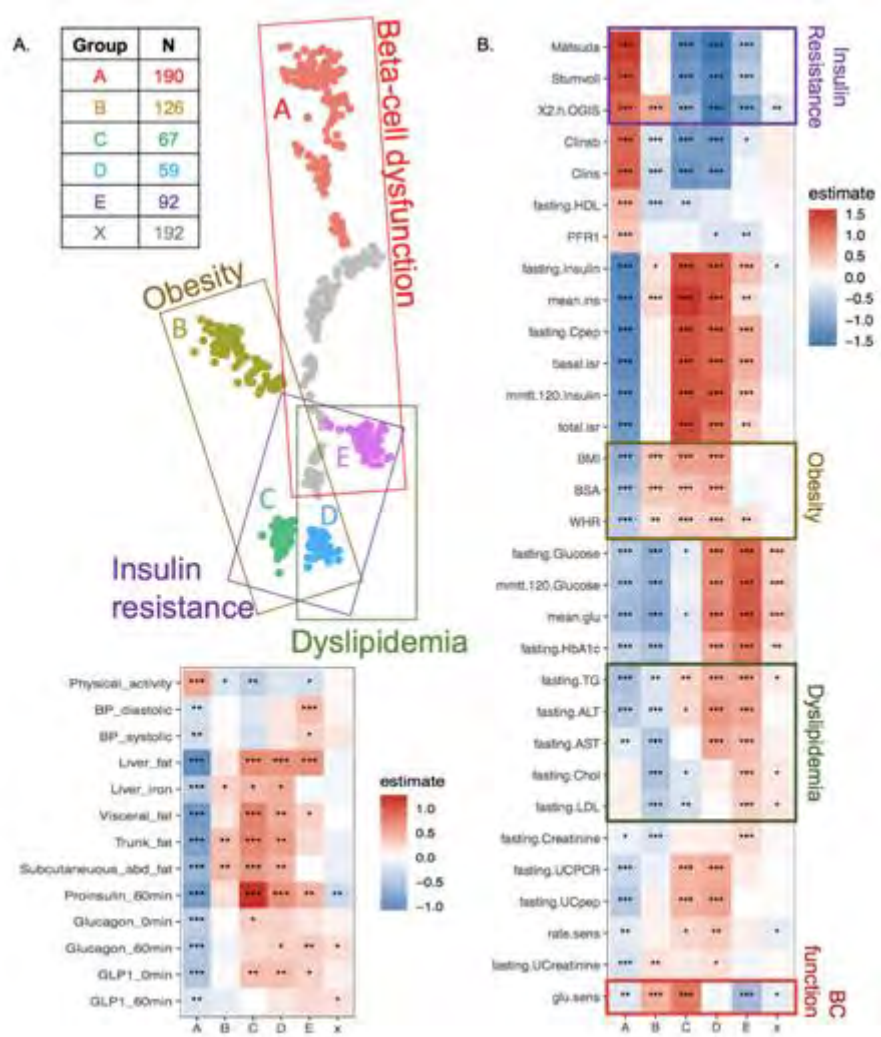


Fig 7: Clinical characteristics of five patient subgroups identified at the extremes of the baseline phenotype spectrum through reversed graph embedding.

- DDRTree representation of the baseline phenotype spectrum of newly-diagnosed T2D patients. Each point represents an individual, and the 5 terminal tree branches are coloured and marked as subgroups A-E. The individuals shown in grey represent the middle of the phenotypic distribution and are not a focus of further analyses in this study. The box overlays schematically illustrate individuals affected with dysfunctions in 4 main T2D aetiological processes: beta-cell dysfunction, obesity, insulin resistance, and dyslipidemia.
- Summary of the 31 clinical variables used in characterization of the baseline T2D phenotypic space. All variables were rank-normally transformed, and for each subgroup and each variable, the heatmap shows the significance level of the Mann-Whitney U test for this subgroup against the remaining individuals from the study. Colour of the heatmap reflects the directionality and magnitude of the test estimate, with red indicating higher, and blue lower values characteristic of the given subgroup.

1.5.4.4.3 Large-scale analysis of whole blood transcriptomic modules in the context of metabolic clinical traits and type 2 diabetes

We used transcriptomics data from 2,127 non-diabetic individuals and 789 T2D patients to describe the co-expression structure of whole blood that mainly reflects processes and cell types of the immune system, and how it relates to metabolically relevant clinical traits and T2D. We identified a large number of associations between transcriptomic modules and measures of insulin sensitivity and glucose tolerance, pointing to a major overlap of immune and metabolic processes (*see figure 8 below*). Some of the metabolically-linked modules reflect neutrophil lymphocyte ratio in blood while others are independent of white blood cell estimates, including a module of genes encoding neutrophil granule proteins (NGPs) with antibacterial properties for which the strongest associations with clinical traits and T2D status were observed. Yet, integration of genetic data did not support a causal role of this module in the development of T2D, but instead highlighted four additional modules as potentially causal that are involved in pathways such as cholesterol homeostasis, glycolysis/gluconeogenesis and semaphoring interactions. Our results provide a large-scale map of whole blood transcriptomic modules in the context of metabolic disease and point to novel biological candidates for future studies.

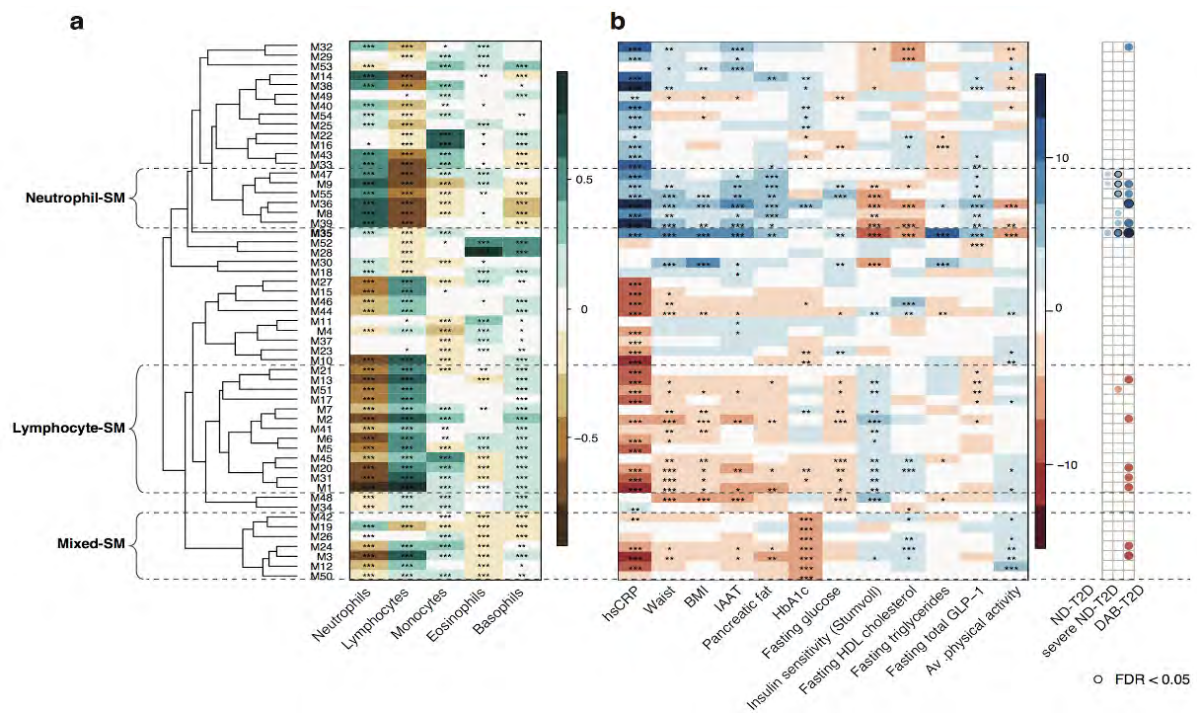


Fig 8: Whole blood transcriptomic co-expression modules show extensive correlations with clinical traits in 2,127 non-diabetic individuals. **a)** Heatmap showing Pearson's correlation between modules (rows) and white blood cell estimates. Stars indicate statistical significance as such: ***FDR<0.001, **FDR<0.01, *FDR<0.05. **b)** Heatmap providing overview of the associations between modules and selected clinical traits. The heatmap colours denote the linear regression estimates, where all phenotypes have previously been rank normal transformed and residualised for age, gender, centre and technical covariates.

1.5.4.5 Drug response in diabetes

1.5.4.5.1. Sulphonylurea response

Considerable inter-individual variation in glycaemic response to SUs exist in patients with T2D. While 15% of those treated with SU show a marked response, 10-20% do not respond, with intermediate response phenotype for the remaining. Higher baseline HbA1c, male sex, shorter duration of diabetes and lower dose have been previously reported to be associated with better treatment outcomes. Genetics may help to explain part of the remaining variation.

In this meta-GWAS, we have included 5,479 subjects of European ancestry from six different cohorts treated with sulphonylureas as monotherapy or add-on to metformin. Multiple linear regression analyses using HbA1c reduction (HbA1c reduction after one year of SU treatment) as the outcome variable were performed by adding appropriate covariates such as: baseline HbA1c, sex, age at diagnosis, average daily dose and BMI.

SNPs located on chromosome 12 reached the genome-wide significant threshold ($p < 5 \times 10^{-8}$) - *figure 9 below*. These SNPs were located in the intronic region of *SLCO1B1/B3* genes (Figure 2). rs10770791 was the lead SNP at this locus. The minor allele (C) of the most strongly associated SNP, rs10770791, had a frequency of 50% and each copy of the minor allele was associated with a 0.11% lower HbA1c reduction (95% CI [0.07 to 0.15], $p = 4.6 \times 10^{-8}$).

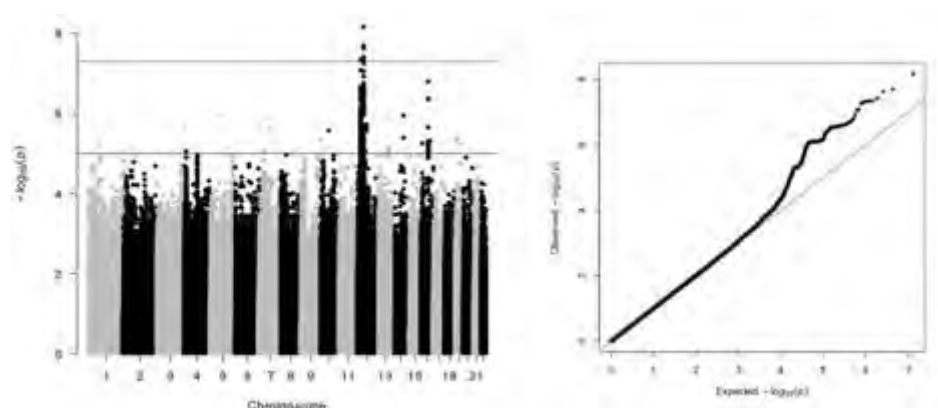


Fig 9: Manhattan and Q-Q plots from the meta-GWAS results of single marker association with glycaemic response to sulphonylureas using an additive genetic model.

Stratified analysis showed a consistent direction of association across all cohorts (Figure 10).

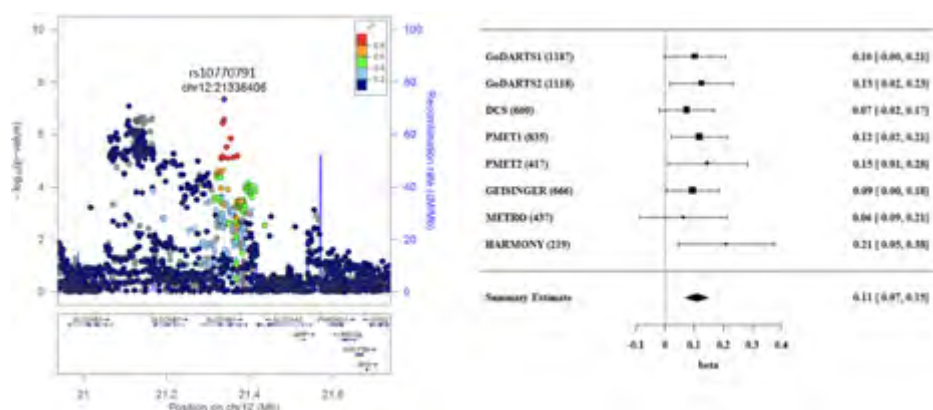


Fig 10: Regional association plot for the *SLCO1B1/B3* and forest plot for the top SNP (rs10770791) across cohorts.

rs10770791 is the top cis eQTL for *SLCO1B1* in the liver tissue (β for C allele = -5.24, $P = 1.61 \times 10^{-7}$) where the C allele was associated with reduced expression. *SLCO1B1* encodes OATP1B1, a protein involved in hepatic uptake of multiple anionic drugs including gliclazide, glibenclamide and glipizide.

This analysis establishes a robust association of an intronic variant in the *SLCO1B1* gene with glycaemic response to sulphonylureas at a genome-wide scale. This gene has strong biological evidence to involve in the pharmacokinetics of sulphonylureas.

1.5.4.5.2 Metformin intolerance

Gastrointestinal adverse effects occur in 20-30% of patients with metformin-treated type 2 diabetes, leading to premature discontinuation in 5-10% of the cases. Gastrointestinal intolerance may reflect localized high concentrations of metformin in the gut. We hypothesized that reduced transport of metformin via the plasma membrane monoamine transporter (PMAT) and organic cation transporter 1 (OCT1) could increase the risk of severe gastrointestinal adverse effects.

The study included 286 severe metformin-intolerant and 1,128 metformin-tolerant individuals from the IMI DIRECT consortium. We assessed the association of patient characteristics, concomitant medication, and the burden of mutations in the *SLC29A4* and *SLC22A1* genes on odds of intolerance. Women ($P < 0.001$) and older people ($P < 0.001$) were more likely to develop metformin intolerance. Concomitant use of transporter-inhibiting drugs increased the odds of intolerance (odds ratio [OR] 1.67, $P < 0.001$) (Figure 1). In an adjusted logistic regression model, the G allele at rs3889348 (*SLC29A4*) was associated with gastrointestinal intolerance (OR 1.34, $P = 0.005$). rs3889348 is the top cis-expression quantitative trait locus for *SLC29A4* in gut tissue where carriers of the G allele had reduced expression. Homozygous carriers of the G allele treated with transporter-inhibiting drugs had more than three times higher odds of intolerance compared with carriers of no G allele and not treated with inhibiting drugs (OR 3.23, $P < 0.001$). Use of a genetic risk score derived from rs3889348 and *SLC22A1* variants found that the odds of intolerance were more than twice as high in individuals who carry three or more risk alleles compared with those carrying none (OR 2.15, $P = 0.01$) (see Figures 11 & 12 below).

These results suggest that intestinal metformin transporters and concomitant medications play an important role in the gastrointestinal adverse effects of metformin.

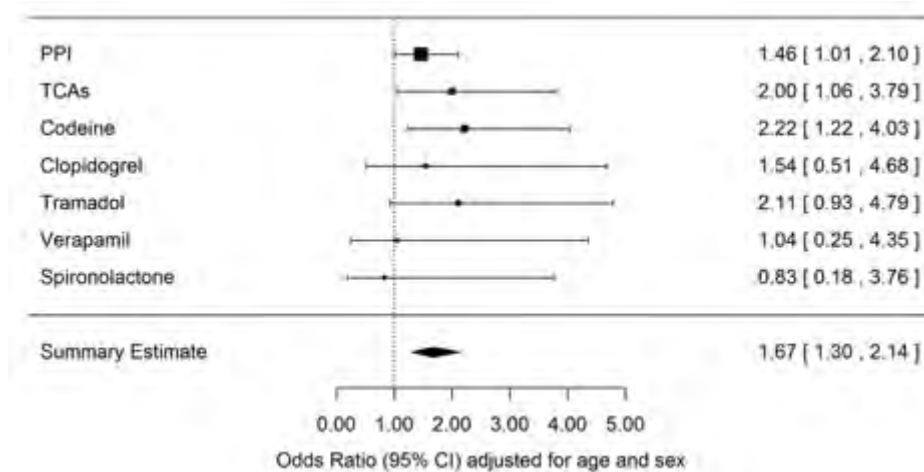


Fig 11: Association of individual intestinal metformin transporter-inhibiting drugs with intolerance.

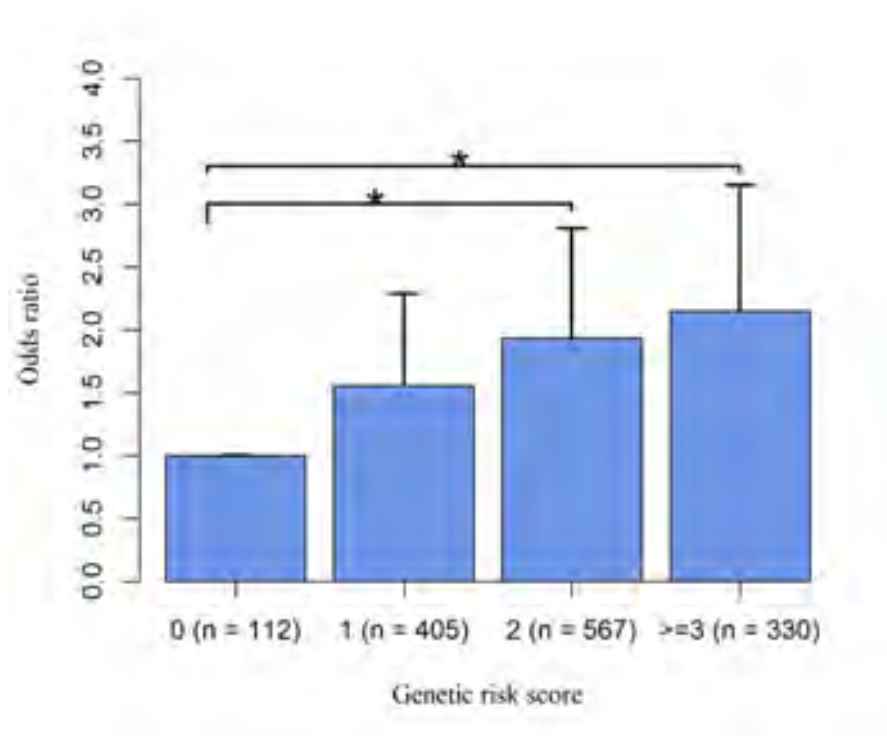


Fig 12: Association of a GRS derived from SLC29A4 (PMAT) and SLC22A1 (OCT1) with metformin intolerance. Bars indicate SE around the mean. *P < 0.05.

1.5.4.5.3 GLP-1RA response

Glycaemic response to GLP-1RA treatment varies markedly among patients with T2D yet the mechanism for this variation is uncertain and it is currently not possible to predict who will respond well and who will respond poorly. Common missense variants in the *GLP-1R* gene have previously been reported to alter GLP-1 mediated insulin secretion in non-diabetic individuals. Low-frequency and rare nonsynonymous variants have also been shown to affect the functionality of the *GLP-1R*. We aimed to investigate the association of candidate variants in the *GLP-1R* gene, GWAS on common

variants and burden test on low-frequency and rare nonsynonymous variants to glycaemic response to GLP-1RA.

We performed a meta-analysis using data from four observational cohorts (DIRECT, PRIBA, GoDARTS and PROMASTER) and two randomized clinical trials from the AWARD and HARMONY. In total 4, 571 T2D subjects were followed-up for 6 months after initiation of GLP-1RA. Association of variants in the *GLP-1R* region, GWAS on common variants and gene based burden test was carried out with reduction in HbA1c after treatment with GLP-1RA using multiple linear regression assuming additive mode of inheritance.

The *GLP-1R* regional association showed independent association of Gly168Ser (rs6923761) and rs2268640 with HbA1c reduction (Gly168Ser β (HbA1c change per allele) = -0.08%, $p = 6 \times 10^{-05}$, rs2268640 β per G allele = -0.10%, $p = 7.02 \times 10^{-06}$). We found no genome-wide significant results that survived multiple testing correction. The top association was for rs61800555 ($b = -0.14$, $p = 2.51 \times 10^{-07}$). rs61800555 is an intronic variant in the *LMX1A* gene. This gene is a transcription factor that binds an A/T-rich sequence in the insulin promoter and stimulates transcription of insulin. The rare variant analysis revealed association of the mutational load of variants in Arrestin, beta1 (*ARRB1*) ($p = 3.94 \times 10^{-07}$). We then derived a genetic risk score by summing up variants in the GLP-1R and *ARRB1* and assessed for association with HbA1c reduction. 5% of the population that carry one or more rare alleles in the *ARRB1* gene had 0.5% (5.5 mmol/mol) greater response compared to 95% of the population that carry no rare allele. *With supporting functional work, this manuscript will be submitted to Nature Genetics in October 2019.*

1.5.5.5 Response to obesity surgery in diabetes

1.5.5.5.1. Multiomic predictors of response to obesity surgery

We have conducted an integrative multiomics analysis of the data available in WP 3.5. The main aim of the project was to identify omics signatures that are important for diabetic remission following bariatric surgery using a forward feature selection setup paired with artificial neural networks. Such omics signatures can both yield insights into remission mechanisms as well as potentially improve prediction of remission.

We have identified 25 omics features from transcriptomics, metabolomics and proteomics data linked to early diabetes remission after RYGB (*see figure 13 below*). Independent verification of the model containing clinical traits alone yielded a predictive accuracy of 0.77, which was further improved ($p < 0.05$) to an accuracy of 0.78-0.85 when adding data on the selected transcripts, metabolites or proteins. This study demonstrates the added value of integrating multiomics with phenotypic data to improve the prediction of diabetes remission after RYGB. *Manuscript in preparation.*

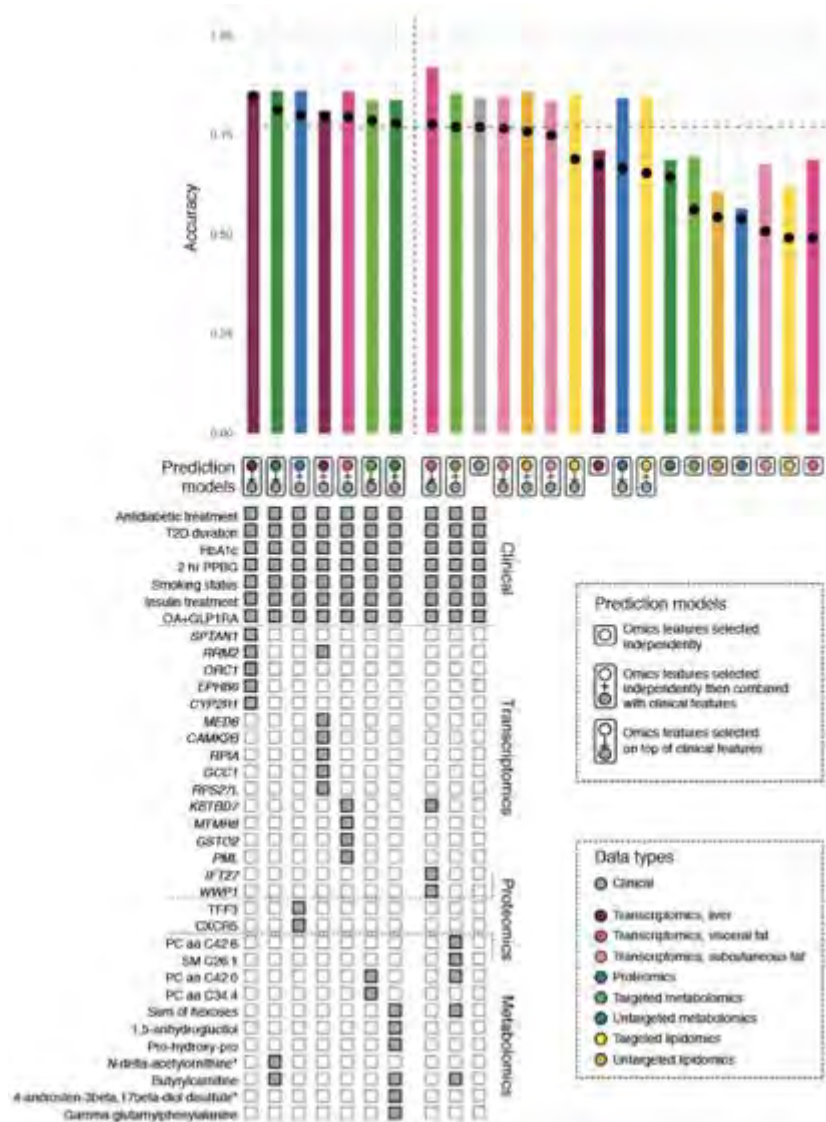


Fig 13: Performance of prediction models. The top panel shows the performance of each prediction model measured by accuracy. Dots represent the performance on the validation data and the bars are sorted according to this measure. Bars are the performance on the discovery data summarized across the five cross-validation folds. Bars are coloured according to the omics data types that are included in the model. The horizontal dotted line, for reference, indicates the performance of the clinical variables alone on the validation data. Of the models performing better than the clinical model the ones to the left of the vertical dotted line had significantly ($p < 0.05$ from bootstrap sampling) higher accuracy on the validation cohort. Features included in the models performing nominally better on the validation data than the purely clinical model are listed in the bottom left panel. Filled boxes indicate the model(s) for which each variable is included. For a full list of selected variables and additional annotation see Table S3. PPBG; Post Prandial Blood Glucose. OA; Oral antidiabetics. GLP1RA; Glucagon-like peptide-1 receptor agonists.

1.6 Potential impact and main dissemination activities and exploitation of results

In 2014, the World Health Organisation estimated that 9% of the world's population had diabetes and over 90% of these suffered from type 2 diabetes. For the European Region, the WHO estimates that there are currently about 60 million people (approx. 10.3% of men and 9.6% of women) aged 25 years and over with diabetes. It is further estimated that half a billion people will suffer from diabetes by 2030. Moreover, type 2 diabetes already causes 5 million deaths per year, mostly from cardiovascular diseases. By 2030, type 2 diabetes is expected to become the 7th leading cause of death globally. Diabetes is also the leading cause of acquired blindness, kidney failure and of lower leg amputations. As a result, global health expenditure on diabetes is estimated at 673 billion dollars (equivalent to 12% of total health expenditure) and this will rise as the pandemic progresses. The total burden of diabetes in Europe is estimated to be around 300 billion €.

Current treatments are far away from an individualized approach; all patients are effectively being treated the same despite the fact that there are twelve different treatment options available. In Western health care systems, diabetes treatment follows an algorithmic sequence that starts with lifestyle modification + metformin (assuming no contraindications). Treatment may progress to other drugs and/or insulin, based upon monitoring tolerability and glycaemic goals. The choice of treatment is often dependent on the stage of the disease, defined by the same surrogate markers, rather than individual patient characteristics. Taking these points into consideration, this leads to differences in efficacy/safety balance in patients and inter-individual variability in responses. Therefore, it is not surprising that about half of diabetes patients do not reach their target goal in HbA1c. Ideally, new approaches would be better tailored to the individual for enhanced tolerability and effectiveness, which would also lead to higher cost effectiveness.

Today, drug development in type 2 diabetic patients is very general, aiming for a reduction in glycated haemoglobin on top of standard treatment. Currently, there are no biomarker(s) predicting, for instance, rapid disease progression, early development of complications, or higher or reduced efficacy. Therefore, large numbers of patients are needed to be enrolled in clinical trials, which makes drug development in type 2 diabetes very expensive. Today, it is estimated that the development of a new anti-diabetic drug will cost around 1.5-2 billion Euro. The lack of predictive biomarkers to identify patients at risk to develop diabetes has also hampered drug development in the early stages of disease, as large sample sizes and long-term trials over many years are required with high risk of failure. Therefore, novel predictive and validated stratification tools will help identify patients at risk to develop diabetes or identify patients with established disease to progress to diabetes complications. In turn, such stratification could allow the design of smaller trials with higher efficiency, a likely higher responder rate and hence, a higher probability of success at reduced costs.

The vast amount of data collected during the 7 years of DIRECT have allowed the largest and most extensively phenotyped database in the world today to be built from prediabetic and T2D patients. The dataset for each individual subject/patient is comprised of phenotypic/clinical, functional and a wide range of 'omics' (proteomics, metabolomics, transcriptomics, genomics, microbiomics) data. The DIRECT database, located in Europe (Denmark), represents a unique foundation for the identification of biomarker(s) that predict patients at risk to develop diabetes, or patients with established disease to

[DIRECT](#)

progress to diabetes complications. In addition, participants' samples are being stored at biobanks (e.g. in Exeter, England), which can be used by internal and external researchers for biomarker discovery and replication. Newly discovered biomarker(s) could provide the basis for a more tailored therapy for patients, with higher efficacy; biomarkers could also be employed prognostically, to allow for smaller clinical trial design with higher efficacy for pharma companies.

A critical step for acceptance and use of any identified biomarker(s) by the scientific/industrial community and healthcare professionals is its validation in prospective clinical studies and subsequent approval by health authorities. Therefore, once biomarker(s) have been identified, engagement with the EMA is planned, to align on the design of a validation trial in order to ensure biomarker acceptance. This will either be done alone or jointly with IMI2 RHAPSODY, as there are overlapping objectives and activities, alongside a number of partners working in both consortia. Of note, IMI2 RHAPSODY has a dedicated work package (WP7) that deals with health authority interactions. An introduction meeting with EMA took place in early 2017, with partners engaged in both DIRECT and RHAPSODY participating.

Any successfully conducted validation trial will be immediately published in a highly ranked clinical journal, as well as being presented at the two largest conferences on diabetes (e.g. Europe: EASD, US: ADA) to ensure a rapid and broad spread of results to the scientific and industrial community, as well as to healthcare professionals.

1.7 Lessons learned and further opportunities for research

The collaborations with EFPIA partners that have occurred during DIRECT have led to the merger of the most effective elements of academic and industrial work cultures. This has been of considerable mutual benefit. At a high level, the academic partners have obtained major insights from EFPIA colleagues concerning operational efficiency, data analysis approaches, and avenues for translating findings (for example, in the design of clinical trials). On the other hand, EFPIA partners have learnt from the academic partners about their challenges in daily patient care leading to specific clinical research questions, insights in way of patient clustering, and the approach to multiomics and clinical data integration. Public-private partnership (PPP) interactions have enabled an exchange of perspectives on project management and structures for executing projects, as well as sharing and presenting the capabilities of novel technologies. From an academic standpoint, continued monitoring of progress towards the project goals – being standard for industry partners – has brought a greater understanding of the different expectations concerning the outcomes of the projects. This has helped to develop a common language and strategy to work towards reaching the milestones and goals.

Most of the output from the project concerns the discovery of biomarkers and complex data analyses that combine biomarkers for the prediction of key outcomes. Thus, the key output has been largely pre-translational. Determining ways in which initial findings can be translated will significantly enhance the value of the project and will represent exciting opportunities for further research.

References for Section 1

1. Shah N, Coathup V, Teare H, Forgie I, Giordano GN, Hansen TH, Groeneveld L, Hudson M, Pearson E, Ruetten H, and Kaye J. Motivations for data sharing-views of research participants from four European countries: A DIRECT study. *Eur J Hum Genet.* 2019 May; 27(5):721-729. doi: 10.1038/s41431-019-0344-2.
2. Shah N, Coathup V, Teare H, Forgie I, Giordano GN, Hansen TH, Groeneveld L, Hudson M, Pearson E, Ruetten H, and Kaye J. Sharing data for future research-engaging participants' views about data governance beyond the original project: a DIRECT Study. *Genet Med.* 2019 May;21(5):1131-1138. doi: 10.1038/s41436-018-0299-7.
3. Fredolini C, Byström S, Sanchez-Rivera L, Ioannou M, Tamburro D, Pontén F, Branca RM, Nilsson P, Lehtiö J, Schwenk JM. Systematic assessment of antibody selectivity in plasma based on a resource of enrichment profiles. *Scientific reports.* 2019 Jun 6;9(1):8324. <https://doi.org/10.1038/s41598-019-43552-5>
4. Häussler RS, Bendes A, Iglesias M, Sanchez-Rivera L, Dodig-Crnković T, Byström S, Fredolini C, Birgersson E, Dale M, Edfors F, Fagerberg L, Rockberg J, Tegel H, Uhlén M, Qundos U and Schwenk J M. Systematic Development of Sandwich Immunoassays for the Plasma Secretome. *Proteomics* 2019 August; 19(15): doi.org/10.1002/pmic.201900008