Diabetologia (2018) 61:138–146
https://doi.org/10.1007/s00125-017-4467-0

ARTICLE

HbA_{1c} is associated with altered expression in blood of cell cycle- and immune response-related genes

Roderick C. Slieker[1,2] · Amber A. W. A. van der Heijden[3] · Nienke van Leeuwen[1] · Hailiang Mei[4] · Giel Nijpels[3] · Joline W. J. Beulens[2,5] · Leen M. 't Hart[1,2,6]

Received: 30 May 2017 / Accepted: 1 September 2017 / Published online: 20 November 2017

Abstract

Aims/hypothesis Individuals with type 2 diabetes are heterogeneous in their glycaemic control as tracked by blood HbA_{1c} levels. Here, we investigated the extent to which gene expression levels in blood reflect current and future HbA_{1c} levels.

Methods HbA_{1c} levels at baseline and 1 and 2 year follow-up were compared with gene expression levels in 391 individuals with type 2 diabetes from the Hoorn Diabetes Care System Cohort (15,564 genes, RNA sequencing). The functions of associated baseline genes were investigated further using pathway enrichment analysis. Using publicly available data, we investigated whether the genes identified are also associated with HbA_{1c} in the target tissues, muscle and pancreas.

Results At baseline, 220 genes (1.4%) were associated with baseline HbA_{1c}. Identified genes were enriched for cell cycle and complement system activation pathways. The association of 15 genes extended to the target tissues, muscle (n = 113) and pancreatic islets (n = 115). At follow-up, expression of 25 genes (0.16%) associated with 1 year HbA_{1c} and nine genes (0.06%) with 2 year HbA_{1c}. Five genes overlapped across all time points, and 18 additional genes between baseline and 1 year follow-up. After adjustment for baseline HbA_{1c}, the number of significant genes at 1 and 2 years markedly decreased, suggesting that gene expression levels in whole blood reflect the current glycaemic state and but not necessarily the future glycaemic state.

Conclusions/interpretation HbA_{1c} levels in individuals with type 2 diabetes are associated with expression levels of genes that link to the cell cycle and complement system activation.

Keywords Blood · Gene expression · Glucose levels · HbA_{1c} · Immune response · RNA sequencing

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00125-017-4467-0) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

✉ Leen M. 't Hart
lmthart@lumc.nl

[1] Department of Molecular Cell Biology, Leiden University Medical Center, Postal Box 9600, 2300 RC Leiden, the Netherlands

[2] Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, the Netherlands

[3] Department of General Practice and Elderly Care Medicine, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, the Netherlands

[4] Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands

[5] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

[6] Molecular Epidemiology Section, Leiden University Medical Center, Leiden, the Netherlands

Abbreviations

DCS	Diabetes Care System
e(GFR)	Estimated (GFR)
FDR	False-discovery rate
GEO	Gene Expression Omnibus
PBMC	Peripheral blood mononuclear cell

Introduction

Individuals with type 2 diabetes are heterogeneous in their disease trajectory, glycaemic control over time [1], response to therapy and in the disease-related complications they develop, including micro- and macrovascular complications [2]. Poor

glycaemic control has been associated with a higher incidence of developing microvascular complications [1, 3]. Therefore, individuals with type 2 diabetes would benefit from new markers for future glycaemic control, especially when in an early stage of the disease.

Much effort has been spent identifying common gene variants that mark disease risk and progression, but genetic variants contribute little in addition to classic risk factors, especially in people below 50 years of age [4]. In addition, genetic risk scores explain only 10–15% of the heritability of type 2 diabetes [5]. Accelerated by recent technological advances, other molecular variables, such as epigenetic modifications and gene expression, are increasingly being investigated in relation to blood glucose and type 2 diabetes and its progression. For example, DNA methylation near known type 2 diabetes loci (for example, *KLF14*, *ZNF518B*, *INS*) is associated with measures of glucose homeostasis (HbA_{1c}, 2 h insulin) in healthy individuals [6, 7].

At the transcriptional level, early studies have found multiple genes to be differentially expressed between the control group and individuals with (pre)diabetes in target tissues [8–11], and also in blood [12–15]. Using a genome-wide approach in peripheral blood mononuclear cells (PBMCs), genes from the c-Jun N-terminal kinase (JNK) and oxidative phosphorylation pathways were differentially expressed in individuals with and without type 2 diabetes [12]. In addition to case–control designs, a limited number of studies have also investigated links between gene expression in blood and target tissues and glycaemic control and disease-related complications. In PBMCs, the expression of genes encoding TNF-α and IL-6 was elevated in individuals with type 2 diabetes with micro- (n = 29) and macroalbuminuria (n = 31) compared with the control group (n = 22) and individuals with type 2 diabetes and normoalbuminuria (n = 18) [13]. In the same study, *TNF* expression correlated with HbA_{1c} levels [13].

While there are indications that measures of glycaemic control are reflected in molecular measures and blood is an interesting tissue from an etiological perspective, the number of studies that have investigated the relationship between gene expression in blood and disease progression is limited. Those that have been conducted have tended to be small cross-sectional case–control studies. We have investigated the relationship between blood gene expression levels and HbA_{1c} levels in almost 400 individuals with type 2 diabetes selected from the Hoorn Diabetes Care System (DCS) cohort [16].

Methods

Study population Individuals who participated in this study are part of the Hoorn DCS cohort, a prospective cohort of over 12,000 individuals with type 2 diabetes [16]. People visit the DCS annually for routine care and data collection, including anthropometric, fasting glucose, HbA_{1c}, blood lipid and blood pressure measurement and information on medication use. A subset of the individuals in the Hoorn DCS cohort are part of a biobank in which biological material is stored for research purposes. Blood RNA was collected in 2013 and 2014 from 1033 individuals who had participated in the biobank previously, without any specific selection criteria; this group were representative of the individuals who visited DCS in 2013 (ESM Table 1). From this group of 1033, we selected 400 individuals (ESM Table 1) based on the following criteria: age at onset between 40 and 75 years; European descent; diabetes duration less than 10 years; and estimated (e)GFR > 30 ml/min. Untreated individuals were excluded. Each participant gave informed consent and the study was conducted in line with the Declaration of Helsinki.

RNA sequencing Blood for RNA was collected in Tempus tubes (ThermoFisher Scientific, Waltham, MA USA), and RNA was isolated from whole blood using the Direct-zol RNA MiniPrep (Zymo Research, Irvine, CA USA). RNA concentrations were determined using Nanodrop (Nanodrop, Wilmington, DE USA) and, in a subset, RNA integrity was examined using lab-on-a-chip (Agilent, Santa Clara, CA, USA). Whole-genome transcriptome data were generated at the human genotyping facility (HugeF) of the Erasmus Medical Center (the Netherlands, www.glimdna.org). RNA sequencing libraries were generated using the Illumina Truseq v2 library preparation kit (Illumina, San Diego, CA, USA). Libraries were paired-end sequenced (50 bp) using the Illumina Hiseq2000.

Samples (n = 44) with a library size smaller than 30 million reads were re-sequenced and the libraries of the first and second run were combined. Reads passing the chastity filter were combined in sets with Illumina's CASAVA. Raw read quality was assessed using FastQC (v0.10.1) [17]. The adaptors identified by FastQC were clipped using Cutadapt (v1.1) using default settings [18]. To trim low-quality ends of the reads, Sickle (v1.2) was used (minimum length 25, minimum quality 20) [19]. Reads were aligned to the genome using STAR (v2.3.0) [20].

To avoid reference mapping bias, SNPs in the Dutch population (Genome of the Netherlands [GoNL]) with minor allele frequency (MAF) > 0.01 in the reference genome were excluded. Read pairs with eight mismatches at most, mapping to five positions at most, were used. Mapping statistics from the binary alignment map files were acquired through Samtools flagstat (v0.1.19-44428cd). The 5′ and 3′ coverage bias, duplication rate and insert sizes were assessed using Picard tools (v1.86). Gene expression, as read count per gene, was calculated using htseq (v0.6.1p1) with default settings based on Ensembl v71 annotation (corresponding to GENCODE v16) [21]. Gene counts were normalised for GC content and gene length using the R package cqn [22]. To exclude sample mix-ups, genotypes of 50 frequently

occurring SNPs were called and compared with available genotype data. Sex was confirmed using gene expression of *XIST* (chromosome X) and *UTY* (chromosome Y). Genes with ≤5 reads in ≥75% of the samples were discarded, as were genes on the sex chromosomes. The final dataset comprised gene expression levels of 391 individuals comprising 15,564 autosomal genes.

Models with blood HbA_{1c} HbA_{1c} was measured using a turbidimetric inhibition immunoassay (Cobas c501, Roche Diagnostics, Mannheim, Germany). All analyses between gene expression and HbA_{1c} at baseline, and 1 or 2 year follow-up were performed using generalised linear models, implemented in the R package edgeR [23]. HbA_{1c} levels were log transformed as they were not normally distributed. The model was adjusted for sex, age, BMI, blood cell composition, metformin dose, sulfonylurea and/or insulin use and technical covariates, as these are factors known to influence gene expression levels and/or HbA_{1c} levels.

In an extended model, additional factors were added including systolic blood pressure, education level (low, mid, high) and smoking status (non-smoker, former, current). Blood cell counts were determined with a UniCel DxH 800 Coulter Cellular Analysis System (Beckman Coulter) and the FC 500 Series system (Beckman Coulter, Brea, CA, USA). Blood cell fractions were also estimated using the R package wbccPredictor [24]. The imputed cell fractions showed a strong correlation with the measured counts (ESM Fig. 1). Blood cell fractions are strongly correlated with each other; therefore, five principal components were included in the model to adjust for the effect of blood cell composition. To investigate the effect of baseline HbA_{1c} on the association at follow-up, we also added baseline HbA_{1c} to the model for the 1 and 2 follow-up in addition to all the other covariates described above. The effect of medication was assessed by performing the model on metformin users only (n = 252), excluding individuals with other forms of (mono/dual) therapy. In the case of missing data or loss at follow-up, the models were performed only with individuals with complete data. The p values for all generalised linear models of the 15,564 genes (15,564 tests) were false-discovery rate (FDR) adjusted using the Benjamini–Hochberg procedure as implemented in the *p.adjust* function in R. A FDR-adjusted p value below 0.05 was considered significant.

Co-expression networks Co-expressed genes (with expression profiles showing a high correlation, suggesting a functional relationship between the genes) were identified using mixed-model co-expression on log-transformed reads per kilobase million (RPKM) values [25]. Mixed-model co-expression is an R-implemented method that uses Pearson correlation while adjusting for confounding, thereby excluding spurious correlations. The method is described in more detail in Furlotte et al. [25]. Genes were considered co-expressed when the absolute correlation was higher than 0.3 with a p value ≤ 0.001. Clusters within the gene co-expression network (i.e. those with a high number of correlated genes) were identified using Cytoscape v3.4.0. Co-expression of genes was plotted using the R package edgebundleR [26]. Graphs were produced using the R package ggplot2 [27].

Gene set enrichment Genes within the three co-expressed clusters were tested for over-representation in gene sets using the default settings of REACTOME (V61) [28]. Pathways with $p_{FDR} < 0.05$ were considered significant.

eQTLs A public expression Quantitative Trait Locus (eQTL) database was used (www.genenetwork.nl/biosqtlbrowser/, accessed July 2017) to identify SNPs that influence gene expression [29]. Genes were mapped to associating SNP based on the Ensembl gene ID. Diabetes-related traits were obtained from the genome-wide association study (GWAS) catalogue and the MAGIC GWAS [30]. The Venn diagram was created using jVenn.

External data Genes identified at baseline were investigated in the target tissues muscle and pancreas, from two external datasets. The first external dataset consisted of gene expression levels in pancreatic islets measured with the Affymetrix Human Gene 1.0 ST Array (Gene Expression Omnibus [GEO] accession number GSE54279), comprising 113 individuals with HbA_{1c} in the range 23.5–85.8 mmol/mol (4.3–10%) and median 39.9 mmol/mol (5.8%) [31].

The second external dataset consisted of gene expression levels in muscle, accessed with the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (GEO accession number GSE18732), comprising 115 individuals with HbA_{1c} range 33.3–136.1 mmol/mol (5.2–14.6%) and median 39.9 mmol/mol (5.8%) with and without type 2 diabetes [32].

Both datasets included expression at the transcript level rather than the gene level. To make the datasets comparable, the average expression of all transcripts of a gene was calculated for 99 genes that could be retrieved in both datasets (out of the 220 genes, 45%) that were present in both datasets. HbA_{1c} levels were converted to International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) HbA_{1c} levels and log transformed. The associations between HbA_{1c} levels and gene expression in muscle and pancreatic islets were determined using Pearson correlation.

Results

Individual characteristics at baseline and follow-up are given in Table 1. Individuals selected for RNA sequencing were a representative subset of all individuals with blood RNA, the

Table 1 Individual characteristics of the sample of the DCS cohort

Characteristic	Baseline (n = 391)	1 year follow-up (n = 372)	2 year follow-up (n = 362)
Sex (% female)	41.2	41.7	41.2
Metformin use (%)	89.5	87.9	89.2
SU use (%)	11.8	15.9	20.4
Insulin use (%)	12.0	12.6	14.9
Age (years)	64.0 (57.3, 70.0)	64.9 (58.3, 71.0)	66.4 (59.4, 72.0)
Diabetes duration (years)	3.7 (2.1, 5.5)	4.6 (3.1, 6.5)	5.6 (4.2, 7.7)
Glucose (mmol/l)	7.9 (7.2, 9.1)	8.0 (7.0, 9.2)	8.1 (7.2, 9.5)
HbA_{1c} (%)	6.4 (6.0, 7.0)	6.5 (6.1, 7.2)	6.6 (6.2, 7.3)
HbA_{1c} (mmol/mol)	47 (42, 53)	48 (44, 55)	49 (44, 56)
BMI (kg/m^2)	29.5 (26.4, 33.0)	29.3 (26.4, 32.7)	29.2 (26.4, 33.0)
LDL-cholesterol (mmol/l)	2.3 (1.8, 2.9)	2.2 (1.8, 2.8)	2.2 (1.7, 2.9)
HDL-cholesterol (mmol/l)	1.2 (1.0, 1.5)	1.2 (1.0, 1.5)	1.2 (1.0, 1.4)
Triacylglycerol (mmol/l)	1.6 (1.1, 2.2)	1.5 (1.1, 2.2)	1.6 (1.1, 2.2)
Systolic BP (mmHg)	134 (124, 152)	138 (126, 151)	136 (126, 151)
Diastolic BP (mmHg)	80 (75, 85)	80 (74, 85)	79 (74, 85)
eGFR (ml/min)	85.8 (73.4, 98.5)	84.4 (71.3, 95.9)	83.7 (70.8, 95.2)

Data are presented as median (first quartile, third quartile) unless otherwise indicated

SU, sulfonylurea

entire cohort and the biobank subset, as their characteristics were very similar (ESM Table 1). Diabetes duration was one of the selection criteria and this was shorter in the group of individuals with RNA sequencing compared with the entire cohort and the biobank subset (ESM Table 1).

Gene expression levels were tested for an association with HbA_{1c}, a measure of glucose levels over the preceding weeks, at baseline and 1 and 2 year follow-up (Fig. 1a). Of the 15,564 genes that passed quality control, 220 genes (1.4%) were associated ($p_{FDR} \leq 0.05$) with HbA_{1c} levels at baseline with adjustment for covariates. Of these, the majority (183 genes) were upregulated (fold change, 1.05–4.80; ESM Table 2) and 37 genes were downregulated (fold change, 1.05–3.34). Blood cell fractions were both measured and estimated based on the gene expression data, but there was no difference in the magnitude of the effect with measurements vs estimates (ESM Fig. 2a). In addition, the observed associations were not driven by differences in medication usage as: (1) all genes showed the same direction of effect in a stratified analysis with metformin users only (n = 252, ESM Fig. 2b); and (2) the effect sizes (log fold change) of the models with and without adjustment for medicine usage were highly correlated (ESM Fig. 2c). To investigate the effect of other factors, such as lifestyle, we extended our model to include systolic blood pressure, education and smoking, but found no difference in the direction or magnitude of effect (r = 0.98, $p < 0.00001$).

The number of genes associated with HbA_{1c} at baseline was considerably higher than at 1 and 2 year follow-up, with 25 genes at 1 year (23 genes upregulated, fold change = 1.17–3.10; two genes downregulated, fold change = 2.33–2.90; ESM Table 2) and nine genes at 2 year follow-up (six genes upregulated, fold change = 1.86–2.69; three genes downregulated, fold change = 1.39–3.88). To identify genes that showed a consistent association with HbA_{1c} over time, the overlap between the three gene sets was determined (baseline and 1 and 2 years; Fig. 1b). Five genes (2.3%) were found to overlap at baseline and both follow-up time points (Fig. 1b,d); 18 additional genes were identified as overlapping at baseline and 1 year follow-up, with no genes overlapping between 1 and 2 year follow-up (Fig. 1b).

As HbA_{1c} levels across years are correlated, particularly between successive years (baseline against 1 year, r = 0.76, and 1 year vs 2 year, r = 0.74; ESM Fig. 3), we next ran the model while adjusting for baseline HbA_{1c}. Of the 25 genes identified, only one remained significantly associated after 1 year follow-up (*NDN,* fold change = −2.11, p_{FDR} = 0.02) and two genes after 2 year follow-up (*FAM132B* [also known as *ERFE*], fold change = −1.90, p_{FDR} = 0.03, and *MTND1P23*, fold change = −3.55, p_{FDR} = 0.04). Moreover, the fold change in the association of genes identified at baseline was largely the same over time without adjustment for baseline HbA_{1c}, but strongly decreased when baseline HbA_{1c} was included in the model (Fig. 1c).

HbA_{1c}-associated genes are involved in the cell cycle and immune response Next, we explored the genes associated with HbA_{1c} at baseline in more detail. First, we investigated whether HbA_{1c} levels causally influence gene expression using Mendelian randomisation. For this, we selected 188 SNPs associated with HbA_{1c} in healthy control individuals

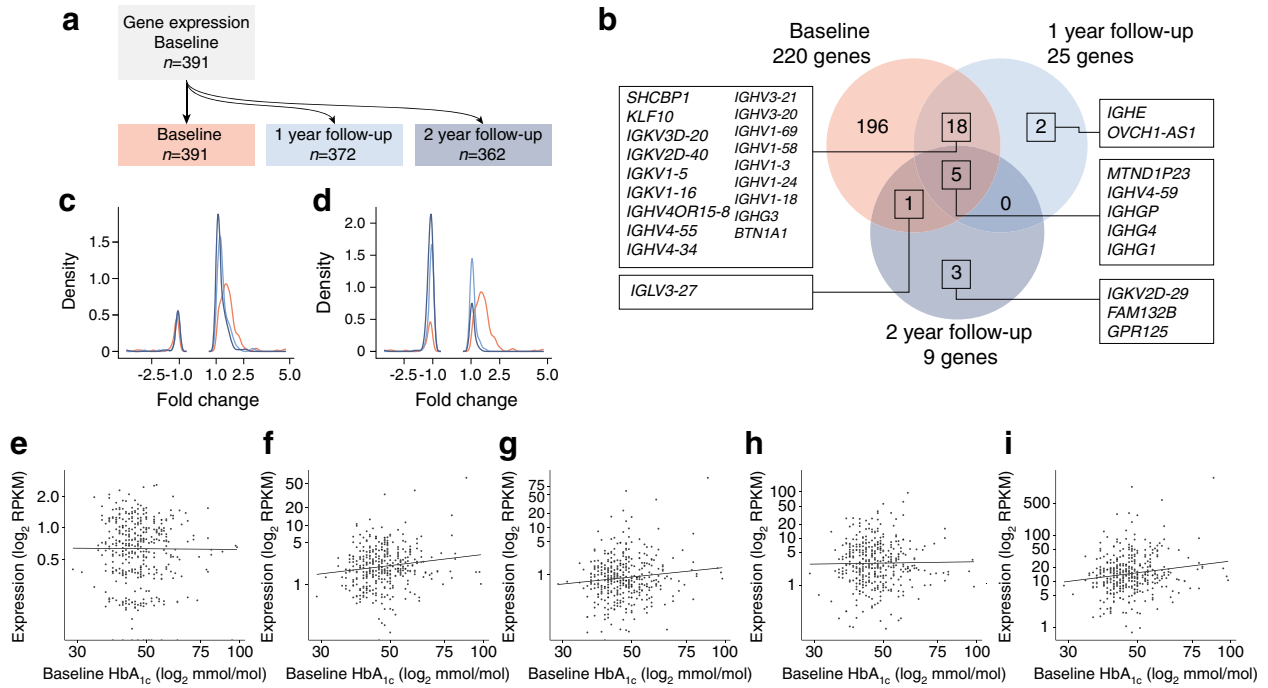

Fig. 1 Association between gene expression levels and HbA_{1c} levels in whole blood. (**a**) Experimental setup. (**b**) Overlap between genes identified as associated with HbA_{1c} at baseline, and at 1 and 2 year follow-up. (**c**, **d**) Density of fold change at baseline (pink), 1 year (light blue) and 2 year (dark blue) follow-up of genes associated with baseline HbA_{1c} levels without adjustment for baseline HbA_{1c} (**c**) or with adjustment for baseline HbA_{1c} (**d**). (**e–i**) Scatterplot of gene expression levels against baseline HbA_{1c} for the five genes identified at each of the follow-up time points: *MTND1P23* (**e**), *IGHV4–59* (**f**), *IGHGP* (**g**), *IGHG4* (**h**) and *IGHG1* (**i**). Data presented are unadjusted for covariates in the model. To convert values for HbA_{1c} in mmol/mol into %, multiply by 0.0915 and add 2.15. GPR125 is also known as *ADGRA3*

from the MAGIC GWAS to serve as genetic instruments [30]. However, when we tested the validity of these genetic instruments in our own data, they did not pass the quality threshold (F value > 10), excluding the possibility of Mendelian randomisation.

Next, we investigated whether identified genes were causally related to the development of type 2 diabetes. Using a public blood eQTL database [29], we identified the 230 strongest associating SNPs near 124 genes (out of the 220). We compared these SNPs to known diabetes-related traits, but found no overlap (ESM Fig. 4). This suggests that there is no relation between known variants involved in type 2 diabetes development and the genes found in this study.

However, several of the 220 genes were found to have a known link to diabetes, including *CD38*, *INSR* and *PC*. *CD38* (fold change = 1.44) is a surface marker associated with insulin resistance in diabetes via the release of inflammatory cytokines [33]. *INSR* (fold change = −1.13, p_{FDR} = 0.03) encodes the insulin receptor important for insulin action. *PC* (fold change = 1.30, $p_{FDR} = 5.1 \times 10^{-3}$) encodes pyruvate carboxylase, which is involved in gluconeogenesis. To more systematically explore the relation between the genes identified, we determined whether they are co-expressed, i.e. whether they are correlated, suggesting a functional link (mixed-model co-expression, $|r| \geq 0.3$, $p \leq 0.001$). Co-expression was found for 99 genes (Fig. 2a), among which three clusters could be distinguished: the largest comprising 55 genes; a second smaller cluster comprising 42 genes; and the third consisting of two genes. The largest cluster showed strong over-representation in cell cycle (checkpoint) pathways (33 genes [15.0%], $p_{FDR} = 1.33 \times 10^{-14}$; Fig. 2 and Table 2). The second cluster showed over-representation for complement system activation and B cell signalling pathway (29 genes [13.2%], $p_{FDR} = 1.11 \times 10^{-16}$; Fig. 2 and Table 2), in line with the large number of genes identified that encode immunoglobulin constituents. The third cluster comprised *KLF10* and *KLF11*, which both link to cell cycle regulation.

Expression of a subset of genes in muscle and pancreas also associates with HbA_{1c} To investigate whether the association with HbA_{1c} extends to target tissues, we expanded the analysis to two external datasets of muscle (n = 115, GSE18732) and pancreatic islets (n = 113, GSE54279) [31, 32]. Of the 220 genes identified at baseline, 99 (45%) were identified in both microarray-based datasets. Of the 37 genes downregulated in blood, several showed a correlation with HbA_{1c} in the same direction ($r < -0.2$, $p \leq 0.05$; Fig. 3a,b) and PAQR7 in both target tissues ($r_{muscle} = -0.31$, $r_{pancreas} = -0.30$, $p < 0.002$; Fig. 3a,b,i–k). Among the 220 upregulated genes in blood, five genes were also found to be upregulated in target tissues: *IGHG1*, *TMEM181* and *RNF19A* in the muscle and *SMC4* and *MCM7* in the pancreas (Fig. 3a,b).

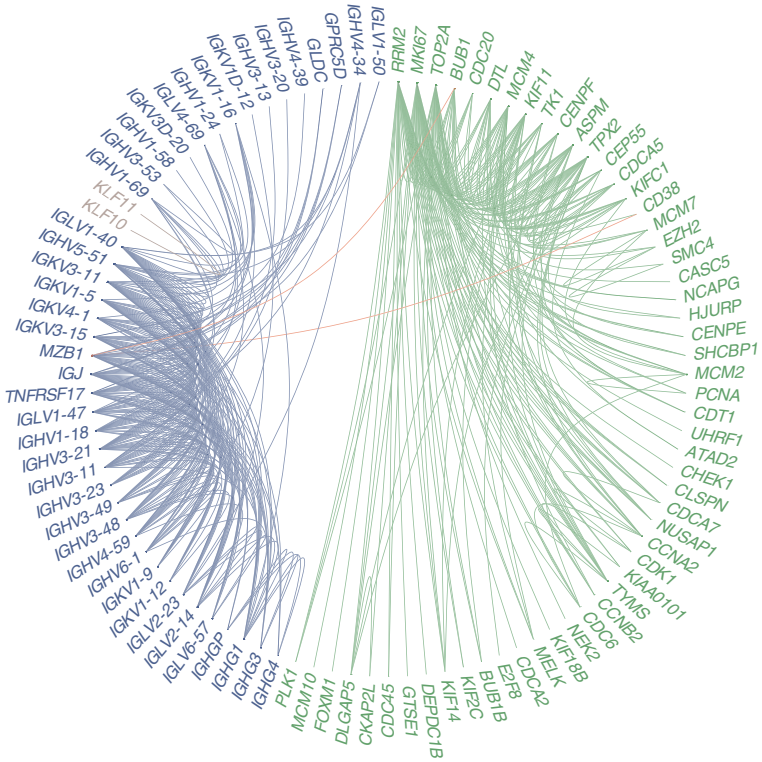

Fig. 2 Co-expression between genes associated with HbA_{1c}. Blue, gene cluster containing immune-related genes; green, gene cluster containing cell cycle-related genes; red, gene co-expression between gene clusters; grey, gene cluster containing two genes from the KLF gene family. *CASC5* is also known as *KNL1*; *KIAA0101* is also known as *PCLAF*; and *IGJ* is also known as *JCHAIN*

Seven genes showed a correlation in the opposite direction ($r < -0.2$, $p \leq 0.02$): *ATAD2*, *CCNF*, *NUF2*, *KIF2C*, *LMAN1*, *GLDC* and *RACGAP1* (Fig. 3a,b). Plots for the five genes showing the strongest correlations in muscle or pancreas are shown in Fig. 3c–q. For muscle, we combined data for individuals with normal glucose tolerance, impaired glucose tolerance and type 2 diabetes. However, when the analysis was performed on individuals with type 2 diabetes only (n = 44, 39%), similar correlations were observed compared with the analysis in all individuals ($r = 0.54$, $p = 4.9 \times 10^{-9}$).

Discussion

In the current study, we investigated the relationship between gene expression levels in whole blood and HbA_{1c} in 391 individuals. The highest number of genes were associated with baseline HbA_{1c}; much lower numbers were associated with HbA_{1c} level at follow-up. The direction of the effect was very similar across the different time points, although a decrease in effect size was observed with time. After adjustment for baseline HbA_{1c}, most correlations of genes with follow-up HbA_{1c}

Table 2 Enrichment of co-expressed gene clusters in REACTOME pathways

Cluster	Pathway identifier	Pathway name	No. genes	No. total	P_{FDR}
1	R-HSA-173623	Classic antibody-mediated complement activation	29	98	1.11×10^{-16}
	R-HSA-2029481	FCGR activation	29	104	1.11×10^{-16}
	R-HSA-5690714	CD22-mediated BCR regulation	22	73	1.11×10^{-16}
	R-HSA-2029485	Role of phospholipids in phagocytosis	29	130	1.11×10^{-16}
	R-HSA-983695	Antigen activates BCR leading to generation of second messengers	22	110	1.11×10^{-16}
2	R-HSA-69278	Cell cycle, mitotic	33	533	1.33×10^{-14}
	R-HSA-1640170	Cell cycle	36	645	1.33×10^{-14}
	R-HSA-453279	Mitotic G1–G1/S phases	14	147	7.13×10^{-13}
	R-HSA-69620	Cell cycle checkpoints	15	188	7.13×10^{-13}
	R-HSA-69206	G1/S transition	13	123	1.22×10^{-12}
	R-HSA-68877	Mitotic prometaphase	13	136	3.57×10^{-12}

FCGR, Fc-gamma receptors; BCR, B cell receptor; no. number; Cluster 1 corresponds to blue genes in Fig. 2; Cluster 2 corresponds to green genes in Fig. 2

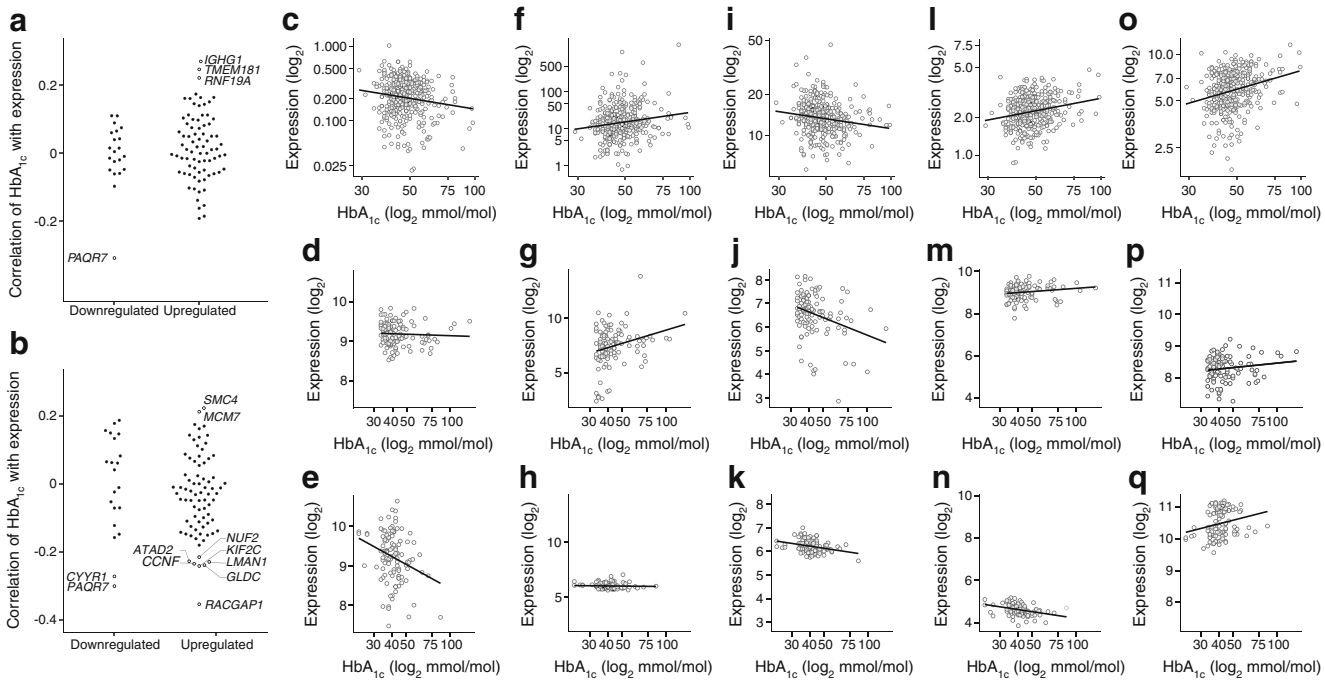

Fig. 3 Association between HbA_{1c} levels and gene expression in muscle and pancreas. (**a**, **b**) Correlation between HbA_{1c} and gene expression for in-blood up- and downregulated genes in muscle. (**c**–**e**) HbA_{1c} against gene expression of *CYYR1*: blood (**c**), fold change = −1.41, $p = 1.7 \times 10^{-2}$; muscle (**d**), $r = -0.05$, $p = 0.60$; pancreas (**e**), $r = -0.27$, $p = 3.7 \times 10^{-3}$. (**f**–**h**) HbA_{1c} against gene expression of *IGHG1*: blood (**f**), fold change = 4.80, $p = 7.25 \times 10^{-9}$; muscle (**g**), $r = 0.27$, $p = 3.6 \times 10^{-3}$; pancreas (**h**) $r = -0.05$, $p = 0.62$. (**i**–**k**) HbA_{1c} against gene expression of *PAQR7*: blood (**i**), fold change = −1.24, $p = 0.01$; muscle (**j**), $r = -0.31$, $p = 8.2 \times 10^{-4}$; pancreas (**k**), $r = -0.30$, $p = 1.3 \times 10^{-3}$. (**l**–**n**) HbA_{1c} against gene expression of *RACGAP1*: blood (**l**), fold change = 1.11 $p = 0.04$; muscle (**m**), $r = 0.17$, $p = 6.4 \times 10^{-2}$; pancreas (**n**), $r = -0.35$, $p = 1.3 \times 10^{-4}$. (**o**–**q**) HbA_{1c} against gene expression of *SMC4*: blood (**o**), fold change = 1.13, $p = 0.05$; muscle (**p**), $r = 0.14$, $p = 0.13$; pancreas (**q**), $r = 0.22$, $p = 1.8 \times 10^{-2}$. *r*, Pearson's correlation coefficient. To convert values for HbA_{1c} in mmol/mol into %, multiply by 0.0915 and add 2.15

lost significance. Genes identified at baseline were enriched for cell cycle and immune pathways.

Baseline HbA_{1c} was associated with 220 genes, but the number of genes strongly decreased over time, with only nine genes associated with HbA_{1c} at 2 years follow-up. This suggests that some genes reflect the HbA_{1c} levels at baseline, but not necessarily future HbA_{1c}. The diminishing relationship was also seen when the genes associated at baseline with HbA_{1c} were followed across time, as fold change of association decreased with time. Moreover, the association with follow-up HbA_{1c} was driven largely by the correlation between HbA_{1c} levels across time; when adjusted for baseline HbA_{1c}, the number of genes associated with follow-up HbA_{1c} further declined.

Our results give insight into the groups of genes that show aberrant expression with different HbA_{1c} levels. We identified three gene clusters as being differentially expressed: one that linked to cell cycle processes, one to immune response and the third consisted of only *KLF10* and *KLF11*. *KLF11* has been described in type 2 diabetes physiology, but has shown mixed results in GWAS [34–36]. A role for the immune system in type 2 diabetes and obesity is increasingly recognised [37, 38], making blood—in addition to target tissues like pancreas and muscle—a relevant tissue to investigate in diabetes. In healthy individuals, exposure to an OGTT leads to changes in expression of immune-related genes over a 2 h period [39]. Moreover, several blood cell types have been suggested to play a role in, for example, insulin resistance [37, 40, 41]. However, the link between the immune system and type 2 diabetes remains complex and controversial. For example, in a Mendelian randomisation study no causal links were found between IL-1 receptor antagonist (IL-1Ra) or C-reactive protein (CRP) and diabetes-related outcomes [42, 43], while IL-1Ra is associated with 2 h glucose and insulin sensitivity [44].

In case–control studies, it has been shown that type 2 diabetes is associated with altered expression of inflammatory and cell cycle genes [12, 14]. Our study suggests that, in addition to having diabetes, the level of glycaemic control is associated with immune- and cell cycle-related alterations in gene expression. Changes in gene expression in other tissues, such as lymph vessels, have also been identified and point to a role of the immune system in diabetes [45]. We also identified genes that were not only associated in blood with HbA_{1c} but also in the muscle and pancreas. Of the genes inversely associated with HbA_{1c} levels, *PAQR7* was downregulated in all three tissues. *PAQR7* is a progesterone receptor that, when activated, promotes glucose tolerance in the mouse GLUTag cell line [46].

In addition to the immune-related genes, we identified genes related to cell cycle and its checkpoints. Six of the cell cycle genes were also confirmed to have a relationship with HbA_{1c} in the pancreas in the same (i.e. *SMC4* and *MCM7*) or opposite direction (*ATAD2*, *CCNF*, *NUF2* and *KIF2C*). Dysregulation of the cell cycle in pancreas and kidneys has been described and linked to a higher risk of developing type 2 diabetes and complications in rodents [47–49]. In humans, SNPs near the cell cycle genes *CDC123* and *CDKN2A* have been found to be associated with increased susceptibility to type 2 diabetes. This suggests that high blood glucose is associated with dysregulation of the cell cycle not only in the pancreas, but also in other tissues.

A limitation of our study is the relatively heterogeneous population of individuals with type 2 diabetes. Individuals have different diabetes histories and use a variety of drugs, including drugs to control their glucose levels. Yet the heterogeneity of individuals is also part of the question, and a biomarker should be independent of a confounding effect of treatment. As the majority of individuals were taking metformin and this drug is dose-dependently associated with HbA_{1c} [50], we adjusted for the metformin dosage and for use of sulfonylureas and insulin (in addition to classic confounders such as sex, age and BMI). However, while we did not observe an effect for differences in, for example, glucose-lowering medication, education, smoking, blood pressure or BMI, it remains a limitation of our study that there may be other factors related to, for instance, lifestyle and concurrent diseases that may have affected HbA_{1c} and gene expression. A second limitation is that we did not replicate our results in an independent cohort; to confirm the validity of our results, we replicated our findings in two different target tissues (pancreatic islets and muscle).

In our study, we measured the gene expression profile of whole blood. While this is the tissue one would want to identify a biomarker in, it should not be confounded by the composition of blood cell subtypes. To adjust for this confounding effect, we estimated and measured the fraction of the five major cell types in blood and adjusted for these cell fractions in the model.

Altogether, while gene expression levels are interesting blood biomarkers for poor glycaemic control, our study suggests that gene expression levels in whole blood reflect current glycaemic state, but are not necessarily predictive of future glycaemic state. The genes identified provide an important insight into the link between poor glycaemic control and altered expression of cell cycle and immune pathways in blood, which, for some genes, also extends to the target tissues muscle and pancreas.

Acknowledgements We thank M. Nannings (DCS West Friesland, Hoorn, the Netherlands) for the excellent technical assistance and P. van 't Hof (Sequencing Analysis Support Core, Leiden University Medical Center, the Netherlands) for bioinformatics support. Some of the data were presented as an abstract at the 53rd EASD Annual Meeting in 2017.

Data availability The datasets generated in the current study are available from the corresponding author.

Funding This work has been funded by BBMRI-NL (Complementary project, CP2013-69), ZonMW Priority Medicine Elderly (grant 113102006 to LMtH). RCS, LMtH and JWJB received support from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement number 115881 (RHAPSODY). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and European Federation of Pharmaceutical Industries and Associations (EFPIA) and is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097-2. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies.

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

Contribution statement RCS, LMtH, JWJB, NvL and HM designed the study. RCS analysed the data and wrote the draft manuscript. AAWAvdH, GN, JWJB and LMtH acquired all the data within the Hoorn DCS cohort. All authors critically read and revised the manuscript and approved the final version of the manuscript. RCS had full access to the data and is the guarantor of this work.

References

1. Walraven I, Mast MR, Hoekstra T et al (2015) Distinct HbA1c trajectories in a type 2 diabetes cohort. Acta Diabetol 52:267–275
2. Raz I, Riddle MC, Rosenstock J et al (2013) Personalized management of hyperglycemia in type 2 Diabetes. Diabetes Care 36:1779
3. UK Prospective Diabetes Study Group (1998) Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 352:837–853
4. de Miguel-Yanes JM, Shrader P, Pencina MJ et al (2011) Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care 34:121–125
5. McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. N Engl J Med 363:2339–2350
6. Yang BT, Dayeh TA, Kirkpatrick CL et al (2011) Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA1c levels in human pancreatic islets. Diabetologia 54:360–367
7. Bacos K, Gillberg L, Volkov P et al (2016) Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes. Nat Commun 7:11089
8. Kodama K, Horikoshi M, Toda K et al (2012) Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. Proc Natl Acad Sci 109:7049–7054
9. Taneera J, Lang S, Sharma A et al (2012) A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. Cell Metab 16:122–134
10. Chen J, Meng Y, Zhou J et al (2013) Identifying candidate genes for Type 2 Diabetes Mellitus and obesity through gene expression

profiling in multiple tissues or cells. J Diabetes Res. https://doi.org/10.1155/2013/970435
11. Dayeh T, Volkov P, Salo S et al (2014) Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS Genet 10:e1004160
12. Takamura T, Honda M, Sakai Y et al (2007) Gene expression profiles in peripheral blood mononuclear cells reflect the pathophysiology of type 2 diabetes. Biochem Biophys Res Commun 361:379–384
13. Navarro JF, Mora C, Gómez M, Muros M, López-Aguilar C, García J (2008) Influence of renal involvement on peripheral blood mononuclear cell expression behaviour of tumour necrosis factor-α and interleukin-6 in type 2 diabetic patients. Nephrol Dial Transplant 23:919–926
14. van der Pouw Kraan TC, Chen WJ, Bunck MC et al (2015) Metabolic changes in type 2 diabetes are reflected in peripheral blood cells, revealing aberrant cytotoxicity, a viral signature, and hypoxia inducible factor activity. BMC Med Genet 8:1
15. Manoel-Caetano FS, Xavier DJ, Evangelista AF et al (2012) Gene expression profiles displayed by peripheral blood mononuclear cells from patients with type 2 diabetes mellitus focusing on biological processes implicated on the pathogenesis of the disease. Gene 511:151–160
16. van der Heijden AA, Rauh SP, Dekker JM et al (2017) The Hoorn Diabetes Care System (DCS) cohort. A prospective cohort of persons with type 2 diabetes treated in primary care in the Netherlands. BMJ Open e015599:7
17. Andrews S (2010) FastQC: A quality control tool for high throughput sequence data (v0.10.1). R package. Available from www.bioinformatics.babraham.ac.uk/projects/fastqc/
18. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12
19. Joshi N, Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (v1.33). R package. Available from https://github.com/najoshi/sickle
20. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 29:15–21
21. Anders S, Pyl PT, Huber W (2015) HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) 31:166–169
22. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics (Oxford, England) 13:204–216
23. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26:139–140
24. van Iterson M (2016) wbccPredictor: A gene expression or DNA methylation based predictor for white blood cell counts (v1.0.1). R package. Available from https://github.com/mvaniterson/wbccPredictor
25. Furlotte NA, Kang HM, Ye C, Eskin E (2011) Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. Bioinformatics (Oxford, England) 27: i288–i294
26. Bostock M, Patrick E, Tarr G (2016) edgebundleR: Circle Plot with Bundled Edges (v0.1.5). R package. Available from https://github.com/garthtarr/edgebundleR
27. Wickham H (2016) In: Gentleman R, Hornik K, Parmigiani G (eds) ggplot2: elegant graphics for data analysis, 2nd edn. Springer, Cham
28. Croft D, Mundo AF, Haw R et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42:D472–D477
29. Zhernakova DV, Deelen P, Vermaat M et al (2017) Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet 49:139–145
30. Soranzo N, Sanna S, Wheeler E et al (2010) Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. Diabetes 59:3229–3239
31. Krus U, King BC, Nagaraj V et al (2014) The complement inhibitor CD59 regulates insulin secretion by modulating exocytotic events. Cell Metab 19:883–890
32. Gallagher IJ, Scheele C, Keller P et al (2010) Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes. Genome Med 2:9
33. Mallone R, Perin PC (2006) Anti-CD38 autoantibodies in type? diabetes. Diabetes Metab Res Rev 22:284–294
34. Florez JC, Saxena R, Winckler W et al (2006) The Krüppel-like factor 11 (KLF11) Q62R polymorphism is not associated with type 2 diabetes in 8,676 people. Diabetes 55:3620–3624
35. Ma L, Hanson RL, Que LN et al (2008) Association analysis of Kruppel-like factor 11 variants with type 2 diabetes in Pima Indians. J Clin Endocrinol Metab 93:3644–3649
36. Neve B, Fernandez-Zapico ME, Ashkenazi-Katalan V et al (2005) Role of transcription factor KLF11 and its diabetes-associated gene variants in pancreatic beta cell function. Proc Natl Acad Sci U S A 102:4807–4812
37. Donath MY, Shoelson SE (2011) Type 2 diabetes as an inflammatory disease. Nat Rev Immunol 11:98–107
38. Grant RW, Dixit VD (2013) Mechanisms of disease: inflammasome activation and the development of type 2 diabetes. Front Immunol 4:50
39. Choi HJ, Yun HS, Kang HJ et al (2012) Human transcriptome analysis of acute responses to glucose ingestion reveals the role of leukocytes in hyperglycemia-induced inflammation. Physiol Genomics 44:1179–1187
40. Jagannathan M, McDonnell M, Liang Y et al (2010) Toll-like receptors regulate B cell cytokine production in patients with diabetes. Diabetologia 53:1461–1471
41. DeFuria J, Belkina AC, Jagannathan-Bogdan M et al (2013) B cells promote inflammation in obesity and type 2 diabetes through regulation of T-cell function and an inflammatory cytokine profile. Proc Natl Acad Sci 110:5133–5138
42. Interleukin 1 Genetics Consortium (2015) Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. Lancet Diabetes Endocrinol 3: 243–253
43. Brunner EJ, Kivimäki M, Witte DR et al (2008) Inflammation, insulin resistance, and diabetes—Mendelian randomization using CRP haplotypes points upstream. PLoS Med e155:5
44. Herder C, Faerch K, Carstensen-Kirberg M et al (2016) Biomarkers of subclinical inflammation and increases in glycaemia, insulin resistance and beta-cell function in non-diabetic individuals: the Whitehall II study. Eur J Endocrinol 175:367–377
45. Haemmerle M, Keller T, Egger G et al (2013) Enhanced lymph vessel density, remodeling, and inflammation are reflected by gene expression signatures in dermal lymphatic endothelial cells in type 2 diabetes. Diabetes 62:2509–2529
46. Flock GB, Cao X, Maziarz M, Drucker DJ (2013) Activation of enteroendocrine membrane progesterone receptors promotes incretin secretion and improves glucose tolerance in mice. Diabetes 62:283–290
47. Gunasekaran U, Gannon M (2011) Type 2 diabetes and the aging pancreatic beta cell. Aging 3:565–575
48. Keller MP, Choi Y, Wang P et al (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res 18:706–716
49. Wolf G (2000) Cell cycle regulation in diabetic nephropathy. Kidney Int 58:S59–S66
50. Hirst JA, Farmer AJ, Ali R, Roberts NW, Stevens RJ (2012) Quantifying the effect of metformin treatment and dose on glycemic control. Diabetes Care 35:446–454