

Applicant: BAILETTI Diego

## **Relazione Finale del progetto finanziato dalla borsa di studio Fo.Di.Ri.-MSD Italia 2015/2016**

Titolo progetto:

**Ri-Sequenziamento di nuova generazione di *loci* associati al diabete ed a caratteri ad esso correlati in studi sull'intero genoma (GWAS), mediante confronto degli estremi di caratteri quantitativi.**

### **Obiettivo generale**

Obiettivo di questo studio era identificare nuove varianti geniche infrequenti o rare ( $< 5\%$  nella popolazione generale) che influenzino la patogenesi del diabete di tipo 2 (DT2), focalizzando l'analisi sui difetti primari della secrezione di insulina nelle cellule beta pancreatiche.

Per raggiungere tale obiettivo il progetto è stato suddiviso in due fasi.

La prima, comprendente il disegno dello studio, la selezione dei geni e dei soggetti ed il loro sequenziamento di nuova generazione (NGS), è stata terminata ed è oggetto di questa relazione finale; la seconda fase, di analisi e conferma dei dati ottenuti in fase 1, è oggetto di richiesta di una borsa di studio per l'anno 2017.

### **Obiettivo Fase 1:**

Questa prima fase prevedeva il ri-sequenziamento profondo, mirato alle regioni codificanti di 9 geni candidati (CDKAL1, ARAP1, JAZF1, IGF2BP2, KCNQ1, ADCY5, ADAMTS9, NAT2 e GCK) selezionati tra quelli riportati dagli studi GWA come associati alla patologia diabetica con livelli di significatività pari a  $p < 10^{-8}$ . Per aumentare la quota di varianti implicate nel controllo del tratto in esame, questa analisi genetica è stata eseguita in individui selezionati dagli estremi fenotipici della secrezione insulinica, misurata mediante il Disposition Index (DI).

### **Sviluppo della Fase 1**

#### **Disegno dello studio**

In questa prima fase del progetto abbiamo sequenziato i geni candidati selezionati che hanno raggiunto una associazione significativa ( $p \leq 5 \times 10^{-8}$ ) con il DT2 nei GWAS, allo scopo di valutare l'influenza dei polimorfismi a bassa frequenza e/o rari sulla malattia. Questo studio rivolto principalmente verso l'individuazione di varianti infrequenti nella popolazione generale, come proposto da molti e recenti lavori scientifici, consentirebbe di scoprire o di chiarire almeno una parte dell'ereditarietà mancante osservata nel DT2 (Manolio TA, Nature 2009). Considerato che le varianti comuni influenzano il tratto con effetto blando, mentre varianti rarissime o private possono influenzare fortemente la funzione proteica e dunque la patogenesi; si ipotizza che varianti a bassa frequenza veicolino effetti di ampiezza intermedia (Peloso GM, Eur J Hum Genet. 2016, Bonnefond A and Froguel P, Cell Metab. 2015; Auer PL and Lettre G, Genome Medicine 2015,). Infatti attualmente sono noti più di 80 loci genetici associati statisticamente al DT2, ma anche considerate tutte insieme, le varianti scoperte spiegano solo una frazione dell'ereditarietà osservabile nella malattia (Fuchsberger C, Nature 2016). Osservati da un punto di vista biologico una gran parte di questi *loci* contengono geni coinvolti nella maturazione e funzione delle cellule  $\beta$  del pancreas.

Dunque abbiamo scelto di focalizzarci su quei geni principalmente coinvolti nella secrezione di insulina, tra i quali: CDKAL1, ARAP1, JAZF1, IGF2BP2, KCNQ1, ADCY5, ADAMTS9, NAT2 e GCK. In tutti i geni selezionati, non era stato ancora eseguito deep-resequencing. Di seguito una breve descrizione delle evidenze scientifiche sulle quali si è basata la selezione dei geni:

CDKAL1: espresso nelle isole pancreatiche umane. Studi GWA (Zeggini E, Science 2007, Scott LJ, Science 2007) mostrano un'associazione tra SNPs nell'introne 5 e il DT2 ed una riduzione della funzione delle beta-cellule.

ARAP1 (CENTD2): è risultato significativamente associato con un aumento dei valori di glucosio nel plasma e una diminuzione di rilascio di insulina glucosio-stimolata, suggerendo un'azione sulla funzione delle cellule beta (Nielsen T, Diabetologia 2011).

JAZF1: proteina nucleare (27 kd) con 3 zinc-fingers. Mostra un'associazione con il DT2 con un OR di 1,10 (1,07-1,13)  $p = 5,0 \times 10^{-14}$  nel modello additivo (Zeggini E, Nat Genet 2008). Grarup et al. hanno associato al diabete l'allele T del rs864745, i cui portatori mostrano una diminuzione del rilascio di insulina per ogni allele (Grarup N, Diabetes 2008; Fogarty MP, Diabetes 2013). Beta-cellule di individui diabetici hanno mostrato una ridotta espressione rispetto ai controlli (Marselli L, PLoS One 2010).

IGF2BP2: IGF2 mRNA-binding protein 2 è associato al DT2 (Scott LJ, Science 2007). Marselli et al. hanno osservato una espressione differenziale in cellule beta di soggetti DT2 rispetto ai controlli (Marselli L, PLoS One 2010).

KCNQ1: canale del potassio voltaggio-dipendente espresso in cellule epiteliali, nel pancreas esocrino ed endocrino e nell'intestino tenue (Thevenod F, Am J Physiol Cell Physiol 2002). Numerose varianti associano a ridotta secrezione insulinica durante l'OGTT. In soggetti affetti da DT2, è stato osservato un effetto incretinico ridotto, ridotta concentrazione di GLP-1 in risposta ad un pasto misto, ridotto potere insulinotropico di GLP-1 e perdita quasi completa della secrezione di insulina in fase tardiva in risposta a GIP (Holst JJ, Mol Cell Endocrinol 2009; Mussig K, Diabetes 2009).

ADCY5: della famiglia ciclasasi, responsabili della sintesi di AMPc. Dupuis et al. lo hanno associato all'elevata glicemia a digiuno tramite rs11708067, che comporta un rischio di DT2 con OR = 1,12 (1,09-1,15),  $p = 9,9 \times 10^{-21}$  (Dupuis J, Nat Genet 2010).

ADAMTS9: della famiglia di proteasi ADAMTS zinco-dipendenti. Lo SNP rs4607103 correla con la secrezione di insulina di prima fase glucosio-stimolata (Simonis-Bik AM, Diabetes 2010).

NAT2: Attiva e disattiva farmaci e carcinogeni a base di arilammina ed idrazina. Polimorfismi danno luogo a differenti fenotipi di N-acetilazione (rapido, intermedio e lento). Lo SNP rs1208, è stato associato a diminuita sensibilità insulinica, indipendentemente dal BMI, associato ai tratti correlati all'insulino-resistenza: l'aumentato glucosio a digiuno, l'emoglobina glicosilata (A1C), il colesterolo totale e quello LDL, i trigliceridi e la malattia coronarica (CAD) (Knowles JW, J Clin Invest 2015)

GCK: gluco-esochinasi, fosforila il glucosio permettendone l'ingresso nella maggior parte delle vie metaboliche cellulari. Una delle 3 isoforme è specifica delle beta-cellule. Mutazioni nel gene, nel suo promotore e nella sua metilazione sono state ampiamente collegate a diverse alterazioni metaboliche tra cui il DT2 e tutte le sue complicanze, al metabolismo del glucosio e ad aumento del rischio cardiovascolare (Muller YL, Diabetologia 2014).

Il pannello di *primer* specifici per i geni selezionati è stato disegnato mediante il software on-line DesignStudio (<http://designstudio.illumina.com/>) di Illumina, tenendo conto delle caratteristiche delle regioni target. Questa selezione è stata prima ottimizzata *in silico* poi testata in laboratorio per assicurare la copertura ottimale di tutte le regioni scelte. L'approccio di sequenziamento scelto è stato il TruSeq Custom Amplicon (TSCA) Low Input che prevede la generazione di ampliconi per mezzo di oligonucleotidi specifici e la successiva amplificazione tramite PCR. In questo modo abbiamo creato una libreria di amplificati contenenti le regioni esoniche dei geni selezionati, compresi 15 nucleotidi fiancheggianti ogni lato le regioni di interesse, così da poter analizzare

anche le alterazioni eventualmente presenti nei noti e presunti siti di splicing. Nel processo di preparazione sono aggiunte delle sequenze univoche (index) per rendere identificabili i campioni di partenza nel successivo sequenziamento, progettato per essere eseguito in parallelo su 96 soggetti contemporaneamente.

Il pannello di sequenziamento è stato disegnato dunque sulle regioni codificanti dei 9 geni sopra elencati (ADAMTS9, ADYC5, IGF2BP2, CDKAL1, JAZF1, GCK, NAT2, KCNQ1 e ARAP) per un totale di 29.685 nucleotidi analizzabili per ogni individuo. La copertura media per campione (coverage) osservata nel sequenziamento è stata  $> 150x$ .

Per aumentare la probabilità di rilevare nuove varianti di sequenza che siano funzionali, il disegno dello studio ha coinvolto individui selezionati dagli estremi di una distribuzione di un tratto continuo. In particolare, dal test orale di tolleranza al glucosio (OGTT), condotto su tutti i pazienti reclutati nella coorte, sono stati calcolati come indici di rilascio di insulina glucosio-stimolata l'indice di sensibilità all'insulina (ISI), la risposta insulinica corretta (CIR) ed il disposition index (DI) risultante ( $CIR \times ISI$ ), e sono stati selezionati i soggetti non-diabetici distribuiti nell'estremo superiore e inferiore del 20° percentile dell'indice DI ( $n=383$ ). Il DI è una misura della capacità delle cellule beta di compensare la resistenza all'insulina. Può essere considerato una misura della funzionalità del pancreas nell'individuo sano ed è in grado di prevedere la normale risposta delle cellule beta adeguata per qualsiasi grado di resistenza all'insulina (Bergman RN, Diabetes 2002).

La strategia del ri-sequenziamento profondo congiunta alla selezione degli estremi di un fenotipo, è stato dimostrato essere efficace nel trovare le varianti di sequenza con una MAF  $< 5\%$ , così come in varianti private che possono contribuire a tratti complessi. Caratteristica decisiva delle nostre coorti è l'elevata fidelizzazione, siamo perciò in grado, in caso di scoperta di mutazioni private, di ricontattare i partecipanti dello studio, espandendo l'analisi all'intera famiglia e cercando l'eventuale co-segregazione della mutazione con il tratto, a conferma del ruolo proposto.

## **RISULTATI OTTENUTI IN FASE 1,**

### **ovvero dal ri-sequenziamento dei 9 geni candidati in 383 soggetti “estremi” della secrezione insulinica.**

In estrema sintesi, nella prima fase abbiamo ri-sequenziato le regioni codificanti e parte delle regioni introniche fiancheggianti, in due sottogruppi ciascuno di circa 190 soggetti, selezionati tra quelli con il più alto e il più basso livello di secrezione insulinica, alla ricerca di varianti differenzialmente presenti nei due sottogruppi.

### **Preparazione librerie, sequenziamento (NGS) e manipolazione bioinformatica dei dati:**

In seguito alla selezione dei soggetti e alla valutazione della qualità del materiale genetico in nostro possesso, abbiamo provveduto alla preparazione ed amplificazione delle librerie geniche tramite i reagenti, il protocollo ed il supporto fornito da Illumina. In seguito al controllo qualitativo delle librerie, ho provveduto a preparare un pool equimolare delle librerie da sottoporre a sequenziamento massivo parallelo. Il pannello ed il protocollo è stato infatti progettato per consentire la corsa di 96 campioni contemporaneamente, attraverso l'uso di sequenze univoche per ogni paziente in associazione ai *primers* di selezione ed amplificazione. Dopo il sequenziamento sul sistema MiSeq delle librerie così ottenute, i dati sono stati processati utilizzando sia software integrati che *on-cloud* forniti dalla stessa Illumina, come BaseSpace (<https://basespace.illumina.com>), sia una pipeline personalizzata con software e supporto forniti da CINECA e dal GALAXY project (<https://galaxyproject.org/>).

I dati grezzi risultanti dal sequenziamento in ogni soggetto (fastaq) sono stati sottoposti ad analisi qualitativa, per verificare le caratteristiche delle reazioni di sequenziamento e per valutarne le probabilità di errori nel riconoscimento delle basi. Sono state così eliminate le basi riconosciute con una qualità bassa, ovvero phred score  $< 20$ . Successivamente è avvenuta l'annotazione primaria e secondaria dei dati che permette il riconoscimento delle varianti e l'annotazione delle loro

caratteristiche, tramite la generazione dei file (vcf) individuali. In breve comparando le sequenze individuali con la sequenza di riferimento sono state identificate le varianti individuali che sono state poi oggetto di interrogazione nei principali database on-line al fine di studiarne e riassumerne le caratteristiche, tra cui: la presenza nei database di polimorfismi noti, la frequenza nelle varie popolazioni (MAF), la plausibile funzionalità e l'associazione a tratti già noti, oltre tutte le caratteristiche qualitative derivanti del sequenziamento. I dati individuali così ottenuti sono stati uniformati e riuniti assieme a quelli clinici, permettendo così la costruzione di un database clinico-genetico sul quale basare le successive analisi dei dati globali e le conseguenti analisi statistiche. Le varianti sono state annotate, attraverso i principali programmi di annotazione bioinformatica, come nonsense (stop), missenso (nonsinonime), splicing e sinonime. Attraverso i principali programmi di predizione si è provveduto a creare un punteggio (score) delle mutazioni in base alla coerenza dei diversi strumenti bioinformatici interrogati riguardo la patogenicità o neutralità della variante, e conseguentemente all'analisi della letteratura scientifica in merito, è stata stilata una lista di prioritizzazione in base all'ampiezza e la direzione dell'effetto ipotizzato sulla proteina (Johansson BB, Diabetologia 2017).

### Risultati Preliminari

Analizzando i dati derivanti dal sequenziamento dei 9 geni su 383 soggetti del campione preliminare, selezionati dai due estremi del DI, per un totale di 29.685 nucleotidi ciascuno, sono state evidenziate 1526 varianti globali. Restringendo il campo alle varianti esoniche (n=839), 530 sono risultate non-sinonime e di stop (Loss of Function, LoF), 22 sono risultate essere inserzioni o delezioni (In/Del), 933 sono risultate sinonime o con nessun probabile effetto e 3 con funzione non nota. Delle varianti totali osservate (n=1526) hanno passato i filtri di qualità (Johansen TA et al. J Lipid Res. 2014) più stringenti solo 467. Di queste 237 sono esoniche di cui: 141 LoF, 14 In/Del, 311 sinonime e 2 con funzione ignota. I dati sono stati successivamente analizzati singolarmente per conferma prendendo in considerazione un filtraggio meno stringente: sono state così eliminate dalle analisi solo le varianti che presentavano un'anomala frequenza allelica (LowFreq), ovvero alleli presenti in proporzioni diverse da quelle dell'omozigosi ed eterozigosi. Di queste sono state selezionate solo quelle esoniche (n=380) di cui le LoF erano 222, le In/Del 20, le sinonime 412 e 2 di funzione non nota. La distribuzione delle varianti in base ai due filtri a differente stringenza nei 9 geni è mostrata in tabella 1.

**Tabella 1.**

| GENE          | PASS       |            |          |           |            | NON LowFreq |            |           |           |            |
|---------------|------------|------------|----------|-----------|------------|-------------|------------|-----------|-----------|------------|
|               | var        | NonSin     | Stop     | In/del    | Sin        | N°          | NonSin     | Stop      | In/del    | Sin        |
| ADAMTS9       | 162        | 51         | 5        | 1         | 109        | 220         | 81         | 6         | 3         | 135        |
| ADYC5         | 57         | 15         | 1        | 2         | 40         | 104         | 37         | 2         | 3         | 64         |
| IGF2BP2       | 35         | 6          | -        | 1         | 28         | 53          | 10         | -         | 1         | 42         |
| CDKAL1        | 33         | 10         | -        | 3         | 20         | 49          | 17         | -         | 5         | 27         |
| JAZF1         | 18         | 7          | -        | 1         | 10         | 27          | 14         | 1         | 1         | 12         |
| GCK           | 41         | 4          | -        | 2         | 35         | 41          | 4          | -         | 2         | 35         |
| NAT2          | 11         | 8          | -        | -         | 3          | 20          | 15         | 1         | -         | 5          |
| KCNQ1         | 43         | 13         | -        | 2         | 28         | 66          | 18         | 1         | 3         | 45         |
| ARAP1         | 62         | 21         | -        | 2         | 38         | 76          | 26         | 1         | 2         | 47         |
| <b>totale</b> | <b>462</b> | <b>135</b> | <b>6</b> | <b>14</b> | <b>311</b> | <b>656</b>  | <b>222</b> | <b>12</b> | <b>20</b> | <b>412</b> |

PASS: filtri di qualità più stringente. NON LowFreq: secondo filtraggio nel quale sono state filtrate negativamente solo le varianti a frequenze alleliche anomale. n: numero varianti. NonSin: varianti non-sinonime. Stop: varianti di stop. In/del: Inserzioni e delezioni. Sin: varianti sinonime

È stata poi verificata la distribuzione delle varianti tra gli estremi del DI in valore assoluto ed in percentuale, definendo <20% DI i soggetti che ricadono nel quintile inferiore del DI e >80% DI i soggetti nel quintile superiore, come riassunto nella tabella 2.

**Tabella 2.**

| Gene          | n          | >80%<br>DI | <20%<br>DI | tot        | % var<br>>80%DI | % var<br><20%DI | mut per >80%DI<br>(n=198) | mut per <20%DI<br>(n=185) |
|---------------|------------|------------|------------|------------|-----------------|-----------------|---------------------------|---------------------------|
| ADAMTS9       | 81         | 148        | 131        | 279        | 0.53            | 0.470           | 0.747                     | 0.708                     |
| ADYC5         | 37         | 28         | 17         | 45         | 0.62            | 0.378           | 0.141                     | 0.092                     |
| IGF2BP2       | 10         | 5          | 8          | 13         | 0.38            | 0.615           | 0.025                     | 0.043                     |
| CDKAL1        | 17         | 12         | 9          | 21         | 0.57            | 0.429           | 0.061                     | 0.049                     |
| JAZF1         | 14         | 5          | 9          | 14         | 0.36            | 0.643           | 0.025                     | 0.049                     |
| GCK           | 4          | 1          | 3          | 4          | 0.25            | 0.750           | 0.005                     | 0.016                     |
| NAT2          | 15         | 6          | 10         | 16         | 0.37            | 0.625           | 0.030                     | 0.054                     |
| KCNQ1         | 18         | 12         | 8          | 20         | 0.60            | 0.400           | 0.061                     | 0.043                     |
| ARAP1         | 26         | 14         | 15         | 29         | 0.48            | 0.517           | 0.071                     | 0.081                     |
| <b>totale</b> | <b>222</b> | <b>231</b> | <b>210</b> | <b>441</b> | <b>0.52</b>     | <b>0.48</b>     | <b>1.16</b>               | <b>1.13</b>               |

n: numero varianti. <20%DI/>80%DI soggetti che ricadono rispettivamente nel quintile inferiore e superiore del DI. tot: totale alleli. % var nei <20%DI / >80%DI: percentuale delle varianti espresse rispettivamente nei due estremi inferiore e superiore del DI. mut per <20%DI / >80%DI: densità di varianti nei due estremi del DI.

La stessa distribuzione e relativa percentuale è stata valutata rispetto agli alleli, così da poter testare le associazioni sia con i modelli dominanti che recessivi. I dati sono mostrati in tabella 3 e figura 1.

**Tabella 3.**

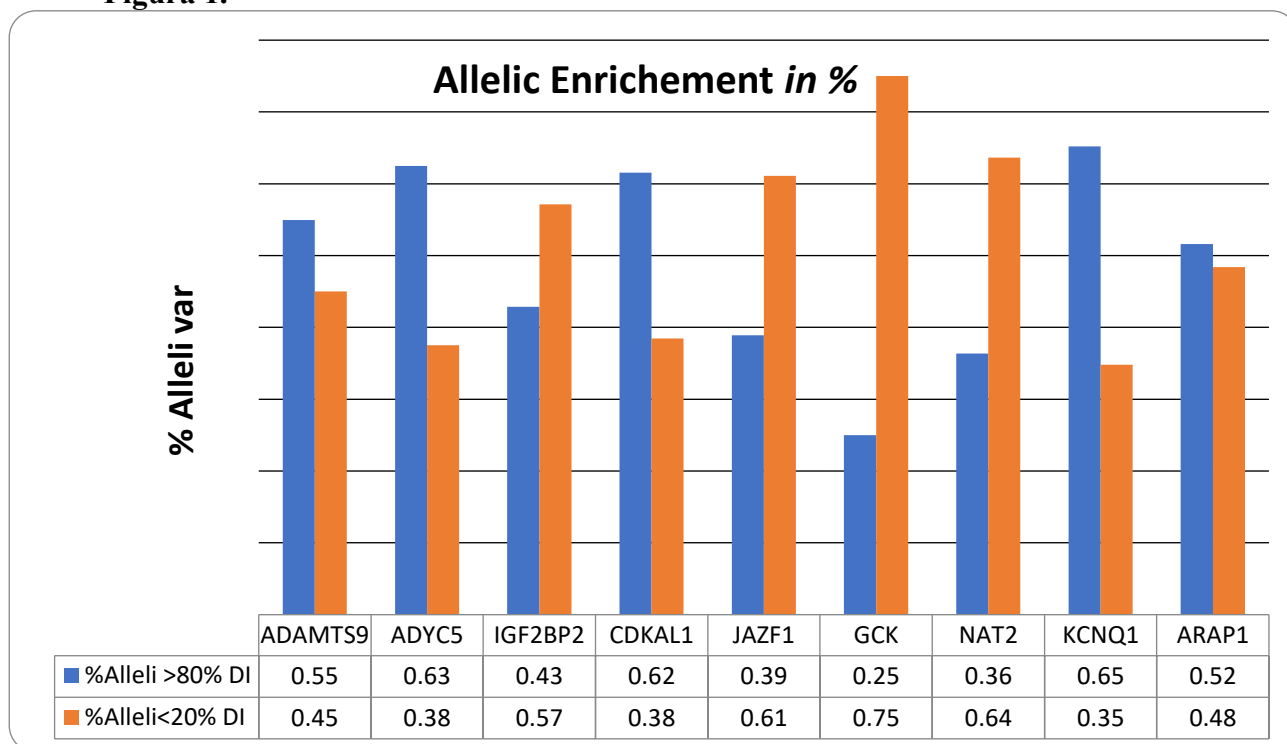
| Gene           | n          | Alleli<br>>80%DI | Alleli<br><20%DI | %Allele<br>>80%DI | %Allele<br><20%DI | p-<br>value | O.R.        |
|----------------|------------|------------------|------------------|-------------------|-------------------|-------------|-------------|
| <b>ADAMTS9</b> | 81         | 199              | 163              | 0.55              | 0.45              | 0.0353      | 1.30        |
| <b>ADYC5</b>   | 37         | 35               | 21               | 0.63              | 0.38              | N.S.        | -           |
| <b>IGF2BP2</b> | 10         | 6                | 8                | 0.43              | 0.57              | N.S.        | -           |
| <b>CDKAL1</b>  | 17         | 16               | 10               | 0.62              | 0.38              | N.S.        | -           |
| <b>JAZF1</b>   | 14         | 7                | 11               | 0.39              | 0.61              | N.S.        | -           |
| <b>GCK</b>     | 4          | 1                | 3                | 0.25              | 0.75              | N.S.        | -           |
| <b>NAT2</b>    | 15         | 8                | 14               | 0.36              | 0.64              | N.S.        | -           |
| <b>KCNQ1</b>   | 18         | 15               | 8                | 0.65              | 0.35              | N.S.        | -           |
| <b>ARAP1</b>   | 26         | 16               | 15               | 0.52              | 0.48              | N.S.        | -           |
| <b>tot</b>     | <b>222</b> | <b>303</b>       | <b>253</b>       | <b>4.39</b>       | <b>4.61</b>       | <b>0.03</b> | <b>1.21</b> |

n: numero varianti osservate nel gene/totali. Alleli<20%DI/>80%DI: numero di alleli variati osservati rispettivamente nell'estremo inferiore e superiore del DI. %Allele <20%DI/>80%DI: percentuale degli alleli variati nei due estremi. p-value e odds ratio (O.R.) sono stati calcolati mediante Chi-Quadro.

Come si osserva è presente una differenza significativa di distribuzione del numero di varianti nel gene ADAMTS9 (Odds Ratio, OR=1.30 p=0.03), ma si osserva anche un trend di distribuzione maggiore verso uno degli estremi negli altri geni.

È stata quindi analizzata la distribuzione globale delle varianti/mutazioni nei due estremi del DI, che è risultata statisticamente significativa con una maggiore frequenza nei soggetti con >80% DI (p<0.03) con un OR di 1.21 (p=0.03). Selezionando solo i 4 geni che mostrano più varianti nel >80%DI (ADAMTS9, ADCY5, CDKAL1 E KCNQ1) l'associazione mostra ampiezza sovrapponibile a quella di ADAMTS9 da solo, aumentando però notevolmente la significatività statistica (OR=1.31 p=0.005).

**Figura 1.**



Distribuzione in percentuale degli alleli nei due estremi nei 9 geni.

Infine una prima analisi statistica dei dati sembra evidenziare che essere portatori di più di due varianti LoF a livello globale in questi 9 geni aumenta la probabilità di essere nell'estremo >80% DI, con una significatività statistica, aggiustando per età, sesso, BMI. In particolare essere portatori di una o più varianti, a livello globale in questi 9 geni sembra determinare un OR rispettivamente di 9.06 (1.73-47.42) con  $p=0.009$  e di 8.52 (1.32-55.01) con  $p=0.024$  (Tabella 4)

**Tabella 4.**

|                               | O.R. (95% C.I.)   | p     |
|-------------------------------|-------------------|-------|
| <b>Mutazioni globali&gt;2</b> | 8.52 (1.32-55.01) | 0.024 |
| <b>Mutazioni globali=1</b>    | 9.06 (1.73-47.42) | 0.009 |

Regressione logistica binaria, corretta per età, sesso e BMI.

È evidente quindi la rilevanza delle osservazioni ottenute in fase 1, dove abbiamo evidenziato un arricchimento di mutazioni associato al >80%DI e che determinano un rischio globale (essendo carrier di >2 varianti) di alterazione nella secrezione di insulina con un OR=8.52 ( $p=0.024$ ). Diventa fondamentale perciò procedere alla fase 2 di analisi e conferma dei dati ottenuti, così da poter aggiungere informazioni sui meccanismi molecolari in cui questi geni sono coinvolti.

Questi dati preliminari sono stati inviati come contributi scientifici ai congressi nazionali ed internazionali SIE 2017 e EASD 2017.

#### **Rilevanza dei risultati ottenuti e di quelli attesi:**

Attraverso questo innovativo approccio al ri-sequenziamento ci proponiamo di analizzare e confrontare soggetti che presentino un tratto fenotipico estremo del carattere quantitativo oggetto dello studio piuttosto che dicotomizzare la stessa distribuzione in casi e controlli, in modo da

massimizzare la potenza statistica dei dati nell'identificare varianti di rischio, o protettive, infrequenti.

Ci aspettiamo che tale studio riesca a fornire molteplici risultati relativi sia alla conferma di associazione del DT2 a geni noti, sia all'individuazione di nuove associazioni con polimorfismi a bassa frequenza e con mutazioni rare, sia a possibili nuove informazioni sui meccanismi biologici che influenzano la patogenesi del DT2. Infatti, è da notare che i *loci* di rischio finora indicati riescono a spiegare solo una piccola parte della ereditabilità del DT2. La nostra ipotesi è che la componente ereditaria non identificabile attraverso gli studi di associazione sull'intero genoma finora condotti (missing heritability) (Manolio TA, Nature 2009), possa essere spiegata dalla presenza di varianti multiple, infrequenti (1-5%) o rare (<1%), che sfuggono alle tecnologie in uso ai GWAS, ma evidenziabili solo mediante il ri-sequenziamento profondo da noi proposto. Già i dati preliminari in nostro possesso sono in linea con questi obiettivi. Ovviamente essendo l'obiettivo del progetto incentrato sulle varianti infrequenti, la numerosità del campione di esplorazione rende le associazioni statistiche, soprattutto per effetti intermedi, molto difficoltose da evidenziare con significatività. Solo testando e analizzando interamente, nella seconda fase, le due ampie casistiche, il potere statistico può aumentare notevolmente e con esso la probabilità di dimostrare e confermare il ruolo delle varianti e dei geni candidati.

#### **PUBBLICAZIONI DERIVATE DA QUESTI STUDI**

1) Bertoccini L, Sentinelli F, Leonetti F, **Bailetti D**, Capoccia D, Cimini FA, Barchetta I, Incani M, Lenzi A, Cossu E, Cavallo MG, Baroni MG. The vitamin D receptor (VDR) functional variant rs2228570 (C>T) does not associate with type 2 diabetes mellitus (T2DM). Endocrine Research. In press 2017.

2) Barchetta I, Capoccia D, Baroni MG, Buzzetti R, Cavallo MG, De Cosmo S, Leonetti F, Leotta S, Morano S, Morviducci L, Prudente S, Pugliese G, Trischitta V; SUMMER Study in Diabetes Group. Collaborators: Bertoccini L, Copetti M, Pibiri C, D'Angelo P, Fallarino M, **Bailetti D**, Alessi E, Basile G. The "Sapienza University Mortality and Morbidity Event Rate (SUMMER) study in diabetes": Study protocol. Nutr Metab Cardiovasc Dis. 2016 Feb;26(2):103-8.

#### **ABSTRACT INVIATI A CONGRESSI:**

1) **European Association for the Study of Diabetes EASD 53° annual meeting 2017**  
**Bailetti D**, Sentinelli F, Bertoccini L, Di Costanzo A, Incani M, Arca M, Leonetti F, Prudente S, Baroni MG.

**Deep re-sequencing of 9 Type 2 Diabetes GWAS loci by comparison of the extremes of dynamic indices of insulin secretion.**

**BACKGROUND AND AIMS:** The susceptibility genes for Type 2 Diabetes (T2D) identified so far are mainly involved in beta-cell function. To this point, Genome-Wide Association Studies (GWAS) have identified a large number of *loci*. Despite this large number, the major part of T2D inheritance is still uncovered. This missing heritability might be explained by multiple, low-frequency variants that are not captured by GWAS. A powerful approach to highlight causal variants is to deep re-sequence candidate genes. To enhance the probability to highlight causal or protective variants together with the appropriate statistical power, we aimed to identify by Next Generation Sequencing (NGS) low-frequency variants in GWAS *loci* for T2D, searching for primary defects in beta-cell insulin secretion. For this, we applied a two stage study design:

**Stage 1**, deep re-sequencing of coding and flanking regions of 9 candidate genes that reach GWAS significance ( $p < 10^{-8}$ ), in individuals selected from the extremes of insulin secretion, adjusted for insulin resistance, i.e. Disposition Index (DI);

**Stage 2**, confirmation of the association by genotyping the variants differently distributed between the two extremes of insulin secretion in larger and independent groups of Italian adults and children (N=3130).

**MATERIALS AND METHODS:** In a large population, very well characterized from OGTT, measures of insulin secretion and resistance have been calculated, including insulinogenic index (IGI30), ISI (insulin-sensitivity index), and DI (IGI30xISI). NGS was performed on MiSeq system (Illumina) with TruSeq Custom Amplicon approach. Variants are investigated by proper bio-informatics tools. Discovered variants will be genotyped by Real-Time PCR or other suitable methods, such as SNP array or genotyping by sequencing.

**RESULTS:** We sequenced 383 subjects from the discovery sample. Preliminary results show more than 1500 variants in this sample. Bioinformatics tools predict that at least 122 of them may have a functional effect on protein, being missense or nonsense. A small but relevant part of passing-filter variants seems to be newly discovered (no rs) or to presumptively affect protein function. We then searched for a different distribution of all the infrequent variants within the two extremes of DI. Variants in one of the genes, ADAMTS9, were significantly associated with the higher extreme (>80%) of DI distribution (OR=1.30  $p=0.03$ ). A similar trend was observed for four of the 9 candidate genes. Multivariate analyses showed that carriers of one or more variants have an OR = 9.06 (1.73-47.42)  $p=0.009$  and OR=8.52 (1.32-55.01)  $p=0.024$  respectively, to be in the >80% extreme of DI.

**CONCLUSION:** The next stage involves the in-depth analysis of all the variants discovered, evaluating distribution, frequency, and possible function, together with the replication studies in large cohorts to confirm the association with altered insulin secretion. We expect that this study will deliver several results: from confirmatory gene association to newer T2D associated polymorphisms, to possible new insights into potential biological mechanisms influencing T2D pathogenesis.

## 2) **Società Italiana di Endocrinologia SIE 39° Congresso Nazionale 2017**

**Bailetti D, Sentinelli F, Bertocchini L, Di Costanzo A, Arca M, Prudente S, Baroni MG.**

**Deep re-sequencing of 9 Type 2 Diabetes GWAS loci by comparison of the extremes of dynamic indices of insulin secretion (i.e. Disposition Index)**

**BACKGROUND:** The Type 2 Diabetes (T2D) susceptibility genes identified so far are mainly involved in beta-cell function. To this point, Genome-Wide Association Studies (GWAS) have identified a large number of loci, but were focused on simple parameters. Despite the loci discovered, the major part of T2D inheritance is still missing. This missing heritability might be explained by multiple, low-frequency variants that are not captured by GWAS. A powerful approach to highlight causal variants is to deep re-sequence candidate genes in the extremes of phenotypes.

**AIMS:** Identify low-frequency variants in GWAS loci for T2D, searching for primary defects in beta-cell insulin secretion through two stages study design:

Stage 1, deep re-sequencing of coding and flanking regions of 9 candidate genes that reach GWAS significance ( $p < 10^{-8}$ ), in individuals selected from the extremes of insulin secretion, adjusted for insulin resistance, i.e. Disposition Index (DI);

Stage 2, confirmation of the association by genotyping the variants differently distributed between the two extremes of insulin secretion in larger and independent groups of Italian adults and children (N=3130).

**METHODS:** In a large population, very well characterized from OGTT, measures of insulin secretion and resistance have been calculated, including insulinogenic index (IGI30), ISI (insulin-sensitivity index), and DI (IGI30xISI). Next Generation Sequencing (NGS) will be performed on MiSeq system (Illumina). Discovered variants will be genotyped by Real-Time PCR.

**RESULTS:** We sequenced 281 of 384 subjects from the discovery sample. This preliminary data are not fully analyzed yet, but are used to draw initial results. Most of variations are physically within introns, as expected, but 40% fall in exons. Bioinformatics tools predicts that 68.63% of them may have a functional effect on protein, being missense or nonsense, as shown in Table 1. A small but relevant part of passing-filter variants seems to be newly discovered (no rs) or to presumptively affect the protein function being non-synonymous.

We expect that this study will deliver several results: from confirmatory gene association to newer T2D associated polymorphisms, to possible new insights into potential biological mechanisms influencing T2D pathogenesis.

Table.1: function's prediction

| Number of effects by function |       |         |
|-------------------------------|-------|---------|
| Type                          | Count | Percent |
| MISSENSE                      | 884   | 63.46%  |
| NONSENSE                      | 72    | 5.17%   |
| SILENT                        | 437   | 31.37%  |

Missense / Silent ratio: 2.0229

## METODOLOGIA E PROCEDURA SPERIMENTALE ESTESA

### Popolazioni di studio:

#### Popolazione 1 (soggetti adulti italiani)

Sono stati reclutati un totale di 2234 soggetti caucasici (range di BMI 16-73 Kg/m<sup>2</sup>). Tutti gli individui sono stati sottoposti ad una completa anamnesi, esami fisici, test di routine di laboratorio, ed ECG. È stata raccolta la storia familiare di obesità e diabete, le abitudini di vita (fumo e attività fisica), BMI e distribuzione del grasso corporeo valutati mediante la circonferenza vita. Mediante un test standard di tolleranza al glucosio per via orale (OGTT) sono state misurate le curve di glucosio e insulina (a 0, 30, 60, 90 e 120 minuti). Sono stati misurati la pressione sistolica, diastolica e i lipidi plasmatici (colesterolo totale, LDL e HDL, e trigliceridi). Sono stati calcolati l'indice di omeostasi della resistenza all'insulina (HOMA-IR) (Matthews DR, 1985), l'indice di sensibilità all'insulina (ISI) (Matsuda M, Diabetes Care 1999) e gli indici di funzione delle cellule beta, come l'indice della risposta insulinica corretta (CIR) (Herzberg-Schafer SA, PLoS One 2010), e il Disposition Index, cioè la secrezione di insulina aggiustata per la sensibilità all'insulina (CIRxISI) (Bergman RN, Diabetes2002). Molti altri parametri sono inoltre disponibili come la percentuale di massa grassa, calorimetria indiretta, rapporto vita/fianchi, emocromo completo, test di funzionalità epatica (ALT, AST, AF), fibrinogeno, test di funzionalità renale (creatinina, urea), test di funzionamento tiroideo (TSH, FT3, FT4, abTPO), metabolismo basale, trattamenti farmacologici. Su un totale di 2234 soggetti nella nostra coorte, 1473 (66%) sono obesi (range di BMI 30-73 Kg/m<sup>2</sup>), 761 (34%) sono normopeso (range di BMI 16-27 Kg/m<sup>2</sup>), 750 (34%) sono risultati pazienti di nuova diagnosi per il diabete di tipo 2 dopo l'OGTT, e 441 (20%) sono risultati con IFG/IGT dopo l'OGTT.

#### Popolazione 2 (bambini e adolescenti italiani con obesità e peso normale)

Sono stati reclutati un totale di 896 bambini/adolescenti obesi, in collaborazione con l'ambulatorio dell'Unità Endocrino Pediatrica dell'Ospedale Regionale per la Microcitemia di Cagliari. Motivo dell'afferenza alla clinica: peso corporeo in eccesso. I criteri di esclusione dalla coorte erano la

presenza di disturbi endocrini o sindromi genetiche, tra cui l'obesità sindromica. Sono stati presi in considerazione la storia familiare di obesità e diabete e le abitudini di vita (attività fisica).

Tutti sono stati sottoposti a OGTT, eseguito secondo le raccomandazioni cliniche per i bambini (Sacks DB, Clin Chem 2002), tranne quando era presente una glicemia a digiuno  $\geq 7$  mmol/L. I soggetti sono stati classificati secondo i criteri diagnostici ADA. Sono stati misurati in tutti i soggetti colesterolo totale, LDL, HDL, trigliceridi, acido urico, ALT e AST. È stato calcolato l'indice di HOMA-IR (Matthews DR, Diabetologia 1985), gli indici ISI120 e CIR120 (Hanson RL, Am J Epidemiol 2000). In tutti i soggetti è stata valutata la presenza di anticorpi circolanti associati al diabete di tipo 1 autoimmune.

## **Fenotipizzazione dei soggetti di studio**

### **Soggetti adulti**

L'indice di massa corporea (BMI) è stato calcolato come peso (kg) diviso altezza al quadrato ( $m^2$ ). Sono state misurate circonferenza vita e fianchi ed il loro rapporto vita/fianchi. La diagnosi di DT2 si è basata sulla presenza di terapie ipoglicemicizzanti e/o criteri ADA. La diagnosi di ipertensione si è basata sulla presenza di elevata pressione arteriosa sistolica ( $\geq 130$  mm Hg) e/o diastolica ( $\geq 85$  mm Hg) e/o sull'assunzione di farmaci anti-ipertensivi. Parametri biochimici su plasma /siero (glucosio, insulina, il profilo lipidico completo, transaminasi ecc) sono stati misurati nello stesso laboratorio con tecniche standard.

### **Bambini e adolescenti**

L'obesità è stata definita secondo le tabelle italiane di crescita per altezza, peso e BMI in persone di età compresa tra i 2-20 anni (Cacciari E, J Endocrinol Invest 2006), ed è stato derivato l'indice SDS-BMI. Gli stadi puberali di sviluppo sono stati determinati secondo le indicazioni di Tanner (Tanner JM, Nutr Rev 1981) e i bambini sono stati divisi in due gruppi: età prepuberale (stadio di Tanner I) e puberale (stadi di Tanner II-V). Nessuno dei soggetti stava prendendo qualsiasi tipo di farmaco.

### **Stime di rilascio di insulina e di sensibilità all'insulina derivate da un OGTT**

L'indice CIR è stato calcolato come segue:  $100 * Ins_{30} / [Glc_{30} * (Glc_{30} - 3,89)]$ . CIR è considerato la misura più opportuna per valutare geneticamente la disfunzione delle cellule beta (Herzberg-Schafer, PLoS One 2010). L'indice ISI è stato valutato come proposto da Matsuda e DeFronzo (Matsuda M, Diabetes Care 1999):  $10000 / (Glc_0 * Ins_0 * Glc_{mean} * Ins_{mean})^{1/2}$ . Il DI è stato calcolato come il prodotto di ISI e di CIR, per stimare la capacità di secrezione di insulina aggiustata per la resistenza all'insulina (Bergman RN, Diabetes 2002). Nei bambini, il campionamento è stato eseguito solo ai tempi 0 e 120 min, e sono stati calcolati: ISI120:  $104 * / (Ins_{120} * Glc_{120})$  e CIR120:  $Ins_{120} / (Glc_{120} * (Glc_{120} - 3,89))$  (Hanson RL, Am J Epidemiol 2000).

### **Analisi statistiche**

Le variabili categoriche sono state confrontate mediante il Chi-Square, corretto all'occorrenza, o il test esatto di Fisher. Le differenze tra variabili continue sono valutate mediante lo Student t-test a due code e mediante ANOVA. È stata utilizzata la trasformazione logaritmica per normalizzare le distribuzioni delle variabili continue. Per controllare gli effetti modulatori di altri fattori confondenti, sono state eseguite analisi di regressione lineare multivariata e analisi di regressione logistica. Le varianti verranno considerate sia come singole, sia alleliche, che nell'insieme, considerando nei portatori l'effetto congiunto collassato tramite Burden Test e SKAT-O. Verranno testati i diversi modelli di interazione, additivo, moltiplicativo, recessivo e dominante. In generale, i valori di  $p < 0,05$  sono presi come statisticamente significativi. Tutte le analisi statistiche sono state effettuate con i software SPSS 20.0 e R (<https://www.r-project.org/>), tramite i relativi pacchetti statistici idonei, tra cui l'optimized sequence kernel association test (SKAT-O).

**Power Calculation**

Con la strategia del ri-sequenziamento ci proponiamo di trovare varianti di sequenza con MAF compresa tra 0,5-5% così come varianti “private” (osservate in un solo individuo/famiglia) che possano contribuire ad influenzare tratti complessi. È stato stimato (Guey LT, Genet Epidemiol 2011) che un campione selezionato di 200 individui nel decimo percentile superiore e inferiore (totale 400 soggetti) di un carattere quantitativo (Disposition Index nel nostro studio) dà un potere di rilevare varianti a bassa frequenza (1%) superiore all’80%, partendo da una popolazione di circa 1400 soggetti. Per lo studio di replica in fase 2 della relazione genotipo-fenotipo, considerando la dimensione delle nostre coorti di campioni (2153 adulti e 896 bambini), abbiamo calcolato una potenza statistica dell’80% ( $p < 0,001$ ) per rilevare un effetto genetico per le varianti con una MAF pari o superiore a 1% (Guey LT, Genet Epidemiol 2011).

## Bibliografia utilizzata

Auer PL and Lettre G, Rare variant association studies: consideration, challenges and opportunities. *Genome Medicine* 2015, 7:16

Bergman RN, Ader M, Huecking K, Van Citters G. Accurate assessment of beta-cell function: the hyperbolic correction. *Diabetes* 2002, 51 Suppl 1: S212-220.

Bonnefond A and Froguel P, Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab.* 2015 Mar, 21(3):357-68.

Cacciari E, Milani S, Balsamo A, Spada E, Bona G, Cavallo L, Cerutti F, Gargantini L, Greggio N, Tonini G, Cicognani A. Italian cross-sectional growth charts for height, weight and BMI (2 to 20 yr). *J Endocrinol Invest* 2006, 29: 581–593.

Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010, 42:105–116.

Fogarty MP, Panhuis TM, Vadlamudi S, Buchkovich ML, Mohlke KL. Allele-specific transcriptional activity at type 2 diabetes-associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the JAZF1 locus. *Diabetes*. 2013 May, 62(5):1756-62.

Fuchsberger C, Flannick J et al. The genetic architecture of type 2 diabetes. *Nature*. 2016 Aug 4;536(7614):41-7.

Grarup N, Andersen G, Krarup NT, Albrechtsen A, Schmitz O, Jørgensen T, Borch-Johnsen K, Hansen T, Pedersen O. Association testing of novel type 2 diabetes risk alleles in the JAZF1, CDC123/CAMK1D, TSPAN8, THADA, ADAMTS9, and NOTCH2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged Danes. *Diabetes* 2008, 57(9):2534-2540.

Guey LT, Kravic J, Melander O, Burt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B, Nilsson P, Almgren P, Kathiresan S, Groop L, Seymour AB, Altshuler D, Voight BF. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genet Epidemiol.* 2011 May, 35(4):236-46.

Hanson RL, Pratley RE, Bogardus C, Narayan KM, Roumain JM, Imperatore G, Fagot-Campagna A, Pettitt DJ, Bennett PH, Knowler WC. Evaluation of simple indices of insulin sensitivity and insulin secretion for use in epidemiological studies. *Am J Epidemiol* 2000; 151(2):190-198.

Herzberg-Schafer SA, Staiger H, Heni M, Ketterer C, Guthoff M, Kantartzis K, Machicao F, Stefan N, Häring HU, Fritsche A. Evaluation of fasting state-/oral glucose tolerance test-derived measures of insulin release for the detection of genetically impaired  $\beta$ -cell function. *PLoS One* 2010, 2;5(12): e14194.

Holst JJ, Vilsboll T, Deacon CF. The incretin system and its role in type 2 diabetes mellitus. *Mol Cell Endocrinol* 2009, 297:127–136.

Johansson BB, Irgens HU, Molnes J, Sztromwasser P, Aukrust I, Juliusson PB, Søvik O, Levy S, Skrivarhaug T, Joner G, Molven A, Johansson S, Njølstad PR. Targeted next-generation sequencing

- reveals MODY in up to 6.5% of antibody-negative diabetes cases listed in the Norwegian Childhood Diabetes Registry. *Diabetologia* 2017, 60:625-35
- Knowles JW, Xie W, Zhang Z, et al. Identification and validation of N-acetyltransferase 2 as an insulin sensitivity gene. *J Clin Invest*. 2015 Apr;125(4):1739-51.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009 Oct 8;461(7265):747-53.
- Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, Marchetti P, Weir GC. Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. *PLoS One* 2010, 13;5(7):e11499.
- Matsuda M, DeFronzo RA. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care* 1999, 22: 1462–1470.
- Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and b-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 1985, 28: 412–419.
- Muller YL, Piaggi P, Hoffman D, Huang K, Gene B, Kobes S, Thearle MS, Knowler WC, Hanson RL, Baier LJ, Bogardus C. Common genetic variation in the glucokinase gene (GCK) is associated with type 2 diabetes and rates of carbohydrate oxidation and energy expenditure. *Diabetologia* 2014 Jul; 57(7):1382-90.
- Mussig K, Staiger H, Machicao F, Kirchhoff K, Guthoff M, Schafer SA, et al. Association of type 2 diabetes candidate polymorphisms in KCNQ1 with incretin and insulin secretion. *Diabetes* 2009;58(7):1715–1720.
- Nielsen T, Sparsø T, Grarup N, Jørgensen T, Pisinger C, Witte DR; Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, Hansen T, Pedersen O. Type 2 diabetes risk allele near CENTD2 is associated with decreased glucose-stimulated insulin release. *Diabetologia* 2011, 54(5):1052-1056.
- Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. *Eur J Hum Genet*. 2016 Jun;24(6):924-30.
- Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. *Clin Chem* 2002; 48: 436–472.
- Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007, 316: 1341-1345.
- Simonis-Bik AM, Nijpels G, van Haften TW, Houwing-Duistermaat JJ, Boomsma DI, Reiling E, van Hove EC, Diamant M, Kramer MH, Heine RJ, Maassen JA, Slagboom PE, Willemsen G, Dekker JM, Eekhoff EM, de Geus EJ, 't Hart LM. Gene variants in the novel type 2 diabetes loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B affect different aspects of pancreatic beta-cell function. *Diabetes* 2010, 59(1):293-301.

Tanner JM, Growth and maturation during adolescence. *Nutr Rev* 1981; 39: 43–55.

Thevenod F, Ion channels in secretory granules of the pancreas and their role in exocytosis and release of secretory proteins. *Am J Physiol Cell Physiol* 2002, 283:C651–C672.

Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007, 316: 1336-1341.

Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008, 40:638–645.