European Journal of Human Genetics (2016) 24, 1337–1343

www.nature.com/ejhg

ARTICLE

Association of the *IGF1* gene with fasting insulin levels

Sara M Willems[1], Belinda K Cornes[2,3], Jennifer A Brody[4], Alanna C Morrison[5], Leonard Lipovich[6,7], Marco Dauriz[2,3,8], Yuning Chen[9], Ching-Ti Liu[9], Denis V Rybin[10], Richard A Gibbs[11], Donna Muzny[11], James S Pankow[12], Bruce M Psaty[13,14], Eric Boerwinkle[5,11], Jerome I Rotter[15], David S Siscovick[16], Ramachandran S Vasan[17,18], Robert C Kaplan[19], Aaron Isaacs[1], Josée Dupuis[9,18], Cornelia M van Duijn[1] and James B Meigs*[,2,3]

Insulin-like growth factor 1 (IGF-I) has been associated with insulin resistance. Genome-wide association studies (GWASs) of fasting insulin (FI) identified single-nucleotide variants (SNVs) near the *IGF1* gene, raising two hypotheses: (1) these associations are mediated by IGF-I levels and (2) these noncoding variants either tag other functional variants in the region or are directly functional. In our study, analyses including 5141 individuals from population-based cohorts suggest that FI associations near *IGF1* are not mediated by IGF-I. Analyses of targeted sequencing data in 3539 individuals reveal a large number of novel rare variants at the *IGF1* locus and show a FI association with a subset of rare nonsynonymous variants ($P_{SKAT} = 5.7 \times 10^{-4}$). Conditional analyses suggest that this association is partly explained by the GWAS signal and the presence of a residual independent rare variant effect ($P_{conditional} = 0.019$). Annotation using ENCODE data suggests that the GWAS variants may have a direct functional role in insulin biology. In conclusion, our study provides insight into variation present at the *IGF1* locus and into the genetic architecture underlying FI levels, suggesting that FI associations of SNVs near *IGF1* are not mediated by IGF-I and suggesting a role for both rare nonsynonymous and common functional variants in insulin biology.
European Journal of Human Genetics (2016) **24,** 1337–1343; doi:10.1038/ejhg.2016.4; published online 10 February 2016

INTRODUCTION

The *IGF1* gene encodes insulin-like growth factor 1 (IGF-I). This hormone has many biological functions involving cell growth, proliferation, and apoptosis.[1] Circulating IGF-I concentrations have been associated with several human diseases, including cardiovascular mortality and cardiovascular risk factors such as age, body mass index, total cholesterol, the presence of diabetes, glomerular filtration rate, and alcohol consumption.[2,3] IGF-I levels are inversely correlated with insulin resistance[3] that may be explained by the insulin-like effects of IGF-I on glucose-uptake. IGF-I is structurally comparable to insulin and they both crossreact with the other's receptor.

Genome-wide association studies (GWASs) of fasting insulin (FI) levels revealed common noncoding single-nucleotide variants (SNVs) near the *IGF1* gene.[4,5] SNV rs35767:A>G (hg18 chr12: g.101399699A>G), located 1.2 kb upstream of *IGF1*, was associated with a 0.010 pmol/l per G allele increase in FI level ($P = 3.3 \times 10^{-8}$) in a large GWAS meta-analysis.[4] Another large GWAS meta-analysis, in largely overlapping samples, revealed rs2114912:G>T (hg18 chr12: g.101453133G>T) as the variant most strongly associated with FI in the *IGF1* region.[5] This variant is located 54.7 kb upstream of the *IGF1* gene and is associated with a 0.024 pmol/l increase in FI per copy of the T allele. These findings have inspired further assessment of the role that the *IGF1* gene plays in insulin biology.

In this paper we hypothesize that the associations of SNVs near the *IGF1* gene with FI (hence insulin resistance) are mediated by circulating IGF-I levels, and that the GWAS variants tag other common or rare functional variants in the *IGF1* region associated with FI levels. To test the first hypothesis, we performed mediation analyses using imputed genotyping array data, and to test the second hypothesis we performed association analyses using deep, high-throughput next-generation targeted sequencing data around *IGF1*. We also examined ENCODE Consortium data sets[6] of regulatory elements by viewing the *IGF1* region in the UCSC Genome Browser[7] in order to generate testable hypotheses about direct functional roles and mechanisms of the noncoding FI-associated GWAS variants.

MATERIALS AND METHODS

An overview of the study design is shown in Supplementary Figure 1.

[1]Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; [2]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA; [3]Department of Medicine, Harvard Medical School, Boston, MA, USA; [4]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA; [5]School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA; [6]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA; [7]Department of Neurology, Wayne State University School of Medicine, Detroit, MI, USA; [8]Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Verona Medical School and Hospital Trust of Verona, Verona, Italy; [9]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA; [10]Boston University Data Coordinating Center, Boston, MA, USA; [11]Human Genome Sequencing Center, Baylor College of Medicine, University of Texas Health Science Center, Houston, TX, USA; [12]Division of Epidemiology and Community Health (J.S.P.), University of Minnesota, Minnesota, MN, USA; [13]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA; [14]Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA; [15]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA, USA; [16]New York Academy of Medicine, New York, NY, USA; [17]Cardiology Section, Department of Preventive Medicine and Epidemiology, Boston University School of Medicine, Boston, MA, USA; [18]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA; [19]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York, NY, USA

*Correspondence: Professor JB Meigs, Massachusetts General Hospital, Division of General Internal Medicine, 50 Staniford Street, 9th Floor, Boston, MA 02114, USA. Tel: +1 617 724 3203; Fax: +1 617 724 3544; E-mail: jmeigs@mgh.harvard.edu

Received 1 July 2015; revised 30 November 2015; accepted 22 December 2015; published online 10 February 2016

Study populations

Individuals of European ancestry from four cohorts of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium were included in this study: the Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study (CHS), Framingham Heart Study (FHS), and the Rotterdam Study (RS).[8]

Mediation cohorts. A total of 5141 nondiabetic individuals of CHS ($n=1717$), FHS ($n=3293$), and RS ($n=140$) were available to contribute to mediation analyses. Genotypic data and both FI and circulating IGF-I levels were available on these participants.

Sequencing cohorts. A total of 3539 nondiabetic individuals (ARIC $n=1761$; CHS $n=967$; and FHS $n=811$) who were part of the CHARGE Targeted Sequencing Study were available for analyses of targeted sequence data with the outcome FI. Of these, 567 of the CHS and 78 of the FHS participants included in these analyses were also included in the mediation analyses. The design of the CHARGE Targeted Sequencing Study, including the cohort sampling design, has been described in detail in Lumley *et al*[9] and Lin *et al.*[10] In summary, to set up the analytic sample a case–cohort design was used in which both a cohort random sample and participants with extreme phenotypes for each of the 14 cardiometabolic traits (atrial fibrillation, blood pressure, BMI, bone mineral density, C-reactive protein, carotid intima–media thickness, echocardiography, electrocardiographic PR and QRS interval, FI, hematocrit, pulmonary function, retinal venule diameter, and stroke) were included. For FI (≥8 h fast), this included a sample of 200 participants (100 ARIC, 50 CHS, and 50 FHS) from the high tail of the distribution in individuals without diabetes, defined as either being diagnosed by a physician (ARIC), treated for diabetes, or having a fasting glucose (FG) >7 mmol/l (ARIC, FHS, and CHS). Three FHS participants with type I diabetes were excluded from selection.

Quantitative trait measurement

FI was measured from fasting plasma (FHS) or fasting serum (CHS and ARIC). In FHS, plasma was collected after a ≥8 h overnight fast and FI was measured on frozen specimen using the DPC Coat-A-Count RIA (total immunoreactive insulin) assay (assay sensitivity 1.2 μU/ml). In CHS (≥12 h fast), FI was measured using a competitive RIA (Diagnostic Products Corp., Malvern, PA, USA). In ARIC (≥8 h fast), FI was measured by radioimmunoassay (125Insulin kit; Cambridge Medical Diagnosis, Bilerica, MA, USA) (assay sensitivity 2 μU/ml). In CHS, circulating IGF-I levels were measured by ELISA (Immunodiagnostic Systems Ltd, Boldon Business Park, Boldon, Tyne & Wear, UK) and in RS by a radioimmunoassay (Medgenix Diagnostics, Brussels, Belgium). BMI was measured using standard methods as previously described.[5]

Genotyping in mediation cohorts

In CHS, genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina (San Diego, CA, USA) 370CNV BeadChip system. The following exclusions were applied: call rate <97%, Hardy–Weinberg equilibrium (HWE) *P*-value $<10^{-5}$, >2 duplicate errors or Mendelian inconsistencies (for reference CEPH trios), heterozygote frequency=0, and SNV not found in HapMap. Samples were excluded from analysis for sex mismatch, discordance with prior genotyping, or call rate <95%. Imputation was performed using BIMBAM v0.99[11] with reference to HapMap CEU using release 22. In the FHS, genotyping was conducted using the Affymetrix (Santa Clara, CA, USA) 500K SNP arrays supplemented with the MIPS 50K array. Samples with call rate <97%, excess Mendelian errors (≥1000), or average heterozygosity outside of 5 SD of mean (<5.758% or >29.958%) were excluded. SNPs with minor allele frequency (MAF) ≥1%, call rate ≥97%, differential missingness *P*-value $\geq 10^{-9}$, and <100 Mendelian errors were used for imputation based on the haplotypes of the HapMap CEU release 22 using the MaCH[12] software. In the Rotterdam Study, genotyping was performed using 550 and 610K Illumina arrays. Exclusion criteria for individuals were excess autosomal heterozygosity, mismatches between called and phenotypic gender, and outliers identified by an IBS clustering analysis. SNVs were excluded for HWE *P*-value $\leq 10^{-6}$, or SNP call rate ≤98%. Genotypes with MAFs >1% were used for imputation using HapMap CEU release 22 as a reference panel. Imputation was performed using MaCH.[12]

Targeted next-generation deep sequencing

Target selection in the CHARGE Targeted Sequencing Study included regions that had been associated with one of 14 cardiometabolic traits by previous GWASs and regions that had been shown to exhibit pleiotropy, and included the *IGF1* gene.[10] Four regions in or near the *IGF1* gene were sequenced at a mean depth of 50×, including 1 kb downstream, all five exons plus flanking regions, and five SNVs upstream that were associated with FI in GWAS:[4,5] rs35767:A>G, rs860598:G>A (hg18 chr12:g.101422576G>A), rs855213: A>G (hg18 chr12:g.101432427A>G), rs35747:G>A (hg18 chr12: g.101436688G>A), and rs2114912:G>T (Supplementary Figure 2). A total of 57.5 kb per copy of the *IGF1* region was sequenced. Sequencing methods were described in detail in Lin *et al.*[10] An extensive quality control (QC) pipeline was implemented, consisting of QC procedures in the sequencing laboratory followed by a series of variant-level filtering steps. These included the exclusion of variants mapping more than 100 base pairs from the requested target capture region, exclusion of variants with a Phred-scaled base quality score[13] <30, with less than two reads of the alternate alleles, and variants with a depth of coverage of <10 total reads. Heterozygote genotypes were removed if their alternate to reference allele ratio was disproportionate (<0.2 or >0.8 for one allele). For strand bias, only variants with alternate allele reads obtained from both the positive and negative strands were kept. Finally, SNPs that had >20% missingness across all samples, more than two observed alleles, or were part of an overly dense SNP cluster (≥3 variants in a 10-bp window) were removed. Using only samples from the cohort random sample subjects, SNPs with HWE *P*-value $<1\times10^{-5}$ were filtered. This criterion was not applied in the samples selected based on extreme phenotypes, potentially enriched for rare variants, to prevent filtering out interesting rare variants with a possible role in disease etiology. To validate sequence-based genotypes, cross-validation was performed with data from the Affymetrix Gene Chip 500K Array Set and 50K Human Gene Focused Panel in 1096 FHS samples. A total of 558 SNPs were shared between the two platforms. After excluding missing genotypes, 98.0% of genotypes were concordant between the two platforms, suggesting high accuracy of the sequenced genotypes. The targeted sequencing data have been submitted to dbGaP (phs000651.v6.p10 (FHS), phs000667.v2.p1 (CHS), and phs000668.v1.p1 (ARIC)).

Variant classification and annotation

Variants identified by sequencing of the *IGF1* locus were classified as common if the MAF was ≥1% and rare if the MAF was <1%. Novel variants were those not found in dbSNP, the 1000 Genomes Project, or ESP 6500 (Exome Sequencing Project).[14,15] Variants were annotated using several bioinformatics sources. ANNOVAR[16] was used to determine whether a variant was synonymous, nonsynonymous, intergenic, upstream (within 1 kb upstream of a transcription start site), downstream (within 1 kb downstream of a transcription end site), intronic, in a 3′ untranslated region (3′UTR), or in a 5′UTR. Variants other than synonymous or nonsynonymous were defined as noncoding. Noncoding variants were predicted to be functional if they were predicted to be highly conserved across species using phastCons,[17] predicted to lie in transcription factor binding sites extracted from the HMR Conserved Transcription Factor Binding Site track of the UCSC Genome Browser,[7] in DNAse I hypersensitive sites or transcription factor binding sites identified by the ENCODE Project,[6] or predicted to be functional using the ORegAnno database.[18] In addition to this functional annotation of the variants present in the targeted sequencing data, we examined GTEx[19] and the ENCODE Consortium regulatory element data sets (including DNAseI hypersensitive sites and histone modifications as well as TFBS Chip-seq) and public transcriptome data in the UCSC Genome Browser to determine whether the known common noncoding FI-associated GWAS variants might be directly functional.

Follow-up genotyping in FHS and lookup of selected rare variants

To verify the influence of variant rs151098426:C>T (hg18 chr12: g.101337467C>T) on FI levels, the variant was genotyped in 1745 FHS offspring and 3372 FHS generation 3 participants with FI levels available that

did not overlap with the FHS participants included in the targeted sequencing analyses. Genotyping was performed using TaqMan (ABI PRISM 7700 HT Sequence Detection System, Applied Biosystems, Foster City, CA, USA) at the Joslin Diabetes Center Advanced Genomics and Genetics Core (Boston, MA, USA). We also did a lookup of the variant in FI exome chip meta-analysis results from the CHARGE diabetes-glycemia working group, including 38 528 samples.

Statistical analyses

All analyses were adjusted for age, sex, BMI, and study design variables (ie, clinic site for CHS and ARIC and recruitment cohort for FHS). FI, in pmol/l, was natural log transformed before analyses to improve normality.

Mediation analyses. To test whether association of FI with GWAS variants in the *IGF1* region (rs35767:A>G, rs860598:G>A, rs855213:A>G, rs35747:G>A, and rs2114912:G>T, pairwise r^2 0.272–1.00 in HapMap2 CEU (see Supplementary Table 1)) is mediated by IGF-I levels, in each cohort (CHS, FHS, and Rotterdam Study) two linear regression models per SNP were fitted, assuming an additive allelic effect. In both models, ln(FI) was the outcome variable. Results from the three cohorts were combined using inverse variance weighted fixed effects meta-analysis as implemented in the R package rmeta.[20] In the first model, age, sex, and BMI were included as covariates, and in the second model IGF-I was added as a covariate. From the models, a ratio $\beta_{SNP\_model2}/\beta_{SNP\_model1} < 1$ would suggest that IGF-I levels explained part of the SNP–FI association.

Analyses of targeted sequence data. The analytic strategy of the targeted sequence data, described briefly below, followed the approach outlined in Lumley *et al*[9] and Lin *et al.*[10]

Four subsets based on functional annotation of rare variants within the *IGF1* locus were tested for association with ln(FI) using the Sequence Kernel Association Test (SKAT).[21] The subsets included (1) nonsynonymous variants, (2) novel nonsynonymous variants, (3) noncoding variants that were predicted to be functional, and (4) novel noncoding variants that were predicted to be functional. FHS used a SKAT test that accounted for family structure.[22] SKAT tests were conducted within the three cohorts (CHS, FHS, and ARIC) and meta-analyzed using a weighted sum of squares of *z*-statistics from single-variant score tests. These variant scores were squared, weighted based on combined allele frequencies across all studies, and summed to create a Q-statistic. The significance of the Q-statistics was determined using an asymptotic distribution, as described in Wu *et al.*[21] The weighted squared *z*-score for each variant divided by the total Q-statistic can be used to identify variants contributing most to the signal. To control type 1 error for this part of the analysis, a *P*-value $<0.05/4=0.0125$ (corrected for four tests: 1 trait×4 subsets of variants) was used to define statistical significance for the SKAT tests.

To test whether rare variant associations were independent of the known FI GWAS hits near the *IGF1* gene, conditional analysis was performed by additionally adjusting for the two common variants rs35767 (FI top hit Dupuis *et al*[4]) and rs2114912 (FI top hit Manning *et al*[5]) (r^2 between these variants = 0.272 in HapMap2 CEU) in the rare variant analysis. As these two variants were not present in the targeted sequence data, rs2162679:C>T (hg18 chr12:g.101395389C>T) was used as a proxy for rs35767:A>G ($r^2=0.915$ in HapMap2 CEU) and rs2607988:G>A (hg18 chr12:g.101454013G>A) was used as a proxy for rs2114912:G>T ($r^2=0.882$ in HapMap2 CEU). Conditional SKAT analyses were performed in each cohort seperately and then meta-analyzed. Similar *P*-values in unconditional and conditional analyses suggest that rare variant associations are independent of the known common variant signals.

Although tests of rare variation were the primary aim of the targeted regional sequencing study, we also tested association of all variants with minor allele count (MAC) ≥ 50 identified by sequencing with ln(FI). In ARIC and CHS, standard additive genetic linear regression models were used, whereas in FHS mixed effects models were used to account for familial correlation. Results from each cohort were meta-analyzed using standard fixed effect inverse variance weighted meta-analysis.[23] *P*-values were obtained from unweighted regression models. Analyses weighted by the inverse of the sampling probability were used to obtain unbiased estimates of effect size.[9] The significance treshold for common variant analyses was set at *P*-value $<1.0\times10^{-3}$ (0.05/49 effective number of independent variants calculated using the Li and Ji approach[24]).

For analyses of follow-up genotyping data in FHS, we used linear mixed effect model to compare the average trait values by genotype category. As we performed two tests (offspring and generation 3 cohorts separately), we considered a *P*-value <0.025 (0.05/2) as significant.

RESULTS

Descriptions of the CHARGE cohort characteristics are depicted in Table 1. Both in the individuals contributing GWAS data and in the targeted sequence samples, women were slightly overrepresented. The mean age ranged from 39 to 71 years in the GWAS samples and from 54 to 72 years in the targeted sequence samples. BMI was in the overweight range in all cohorts. As previously observed, FI values varied widely across studies.[4] The same was observed for the IGF-I levels in the GWAS samples.

Mediation analyses

Mediation analyses results are depicted in Table 2. Meta-analyses *P*-values were nominal to borderline significant for each SNV in both models ($P=0.05$–0.15). However, effect estimates were similar to the effect estimates in up to 51 750 samples in the discovery meta-analysis[5] and in FHS, the largest contributing cohort, *P*-values were nominally significant for each SNV in both models ($P=0.01$–0.04) (Table 2). Both in the meta-analysis and in FHS alone, effect estimates were similar between model 1 (ln(FI) ~ SNP+age+sex+BMI) and model 2 (ln(FI) ~ SNP+age+sex+BMI+IGF-I). This is consistent with an effect of the variants near *IGF1* on FI levels that is not mediated by circulating IGF-I levels.

Table 1 Descriptions of the study populations

	GWAS samples			*Targeted sequence samples*		
	CHS	*FHS*	*RS*	*ARIC*	*CHS*	*FHS*
N (% men)	1717 (36.7)	3293 (47.3)	140 (48.6)	1761 (49.7)	967 (44.7)	811 (48.3)
Age (years)	71.6 (4.8)	39.9 (8.8)	66.2 (5.7)	54.7 (5.7)	72.5 (5.4)	54.1 (10.7)
BMI (kg/m^2)	26.1 (4.3)	27.0 (5.4)	26.4 (4.0)	27.2 (5.7)	26.4 (5.0)	27.9 (6.5)
FI (pmol/l)	72.2 (42.7)	30.9 (20.1)	90.1 (53.0)	83.1 (73.2)	103.1 (63.9)	32.6 (21.3)
IGF1 (ng/ml)	96 (32.7)	131.1 (42.8)	136.7 (53.3)	NA	NA	NA

Abbreviations: ARIC, Atherosclerosis Risk in Communities Study; BMI, body mass index; CHS, Cardiovascular Health Study; FHS, Framingham Heart Study; FI, fasting insulin; IGF1, insulin-like growth factor-1; RS, Rotterdam Study.
Values are mean (SD) unless otherwise indicated.

Table 2 Association of known fasting insulin GWAS SNPs in the *IGF1* region with fasting insulin levels without and with IGF1 levels as covariate in the model

	CHS			*FHS*			*RS*			*Meta*			*Discovery paper*[a]		
	β	*SE*	P	*β*	*SE*	P	*β*	*SE*	P	*β*	*SE*	P	*β*	*SE*	P
Model1: ln(FI) ~ SNP+age+sex+BMI															
rs2114912:G>T	0.020	0.024	0.41	−0.039	0.015	0.01	0.002	0.093	0.98	−0.021	0.013	0.09	−0.024	0.004	3.4×10^{-11}
rs860598:G>A	0.007	0.020	0.72	−0.032	0.014	0.02	−0.072	0.076	0.34	−0.020	0.011	0.07	−0.021	0.003	6.9×10^{-10}
rs35747:G>A	0.005	0.019	0.81	−0.032	0.014	0.02	−0.079	0.079	0.32	−0.021	0.011	0.06	−0.021	0.004	8.9×10^{-10}
rs855213:A>G	0.005	0.020	0.81	−0.032	0.014	0.02	−0.072	0.076	0.34	−0.021	0.011	0.06	−0.021	0.004	1.0×10^{-9}
rs35767:A>G	0.013	0.020	0.50	−0.031	0.015	0.04	−0.127	0.080	0.11	−0.017	0.012	0.15	−0.022	0.004	2.4×10^{-9}
Model2: ln(FI) ~ SNP+age+sex+BMI+IGF1															
rs2114912:G>T	0.018	0.024	0.45	−0.039	0.015	0.01	0.004	0.094	0.97	−0.022	0.013	0.08	NA	NA	NA
rs860598:G>A	0.004	0.020	0.85	−0.032	0.014	0.02	−0.071	0.077	0.36	−0.020	0.011	0.07	NA	NA	NA
rs35747:G>A	0.001	0.019	0.95	−0.033	0.014	0.02	−0.078	0.080	0.33	−0.022	0.011	0.05	NA	NA	NA
rs855213:A>G	0.002	0.020	0.94	−0.032	0.014	0.02	−0.071	0.077	0.36	−0.023	0.011	0.05	NA	NA	NA
rs35767:A>G	0.010	0.020	0.61	−0.031	0.015	0.04	−0.125	0.081	0.12	−0.018	0.012	0.13	NA	NA	NA

Abbreviations: CHS, Cardiovascular Health Study ($n=1717$); FHS, Framingham Heart Study ($n=3293$); RS, Rotterdam Study ($n=140$).
[a]Manning *et al*[5] (*n* up to 51 750).

Table 3 Descriptions of known and novel SNPs in the *IGF1* region in the CHARGE Targeted Sequencing Study cohorts combined

	Known	*Novel*[a]	*Total*
No. of SNPs	248	1145	1393
No. of rare SNPs	133	1143	1276
Coding variants (n = *17*)			
Synonymous	2	4	6
Nonsynonymous	5	6	11
Noncoding variants (n = *1376*)			
Intergenic	165	793	958
Upstream	7	24	31
Downstream	5	20	25
Intronic	39	148	187
UTR3	24	146	170
UTR5	1	4	5
Predicted functional[b]	32	156	188

Values are frequencies.
[a]Not known in dbSNP, 1000 genomes project, or ESP 6500.
[b]Predicted transcription factor binding site (ENCODE ChipSeq, HMR) and/or DNAse hypersensitive site (ENCODE DHS) and/or ORegAnno regulatory variant and/or highly conserved (PhastCons).

Table 4 SKAT meta-analyses results for fasting insulin (BMI adjusted) from different subsets of rare (MAF <1%) SNPs in the *IGF1* region

Subset of rare SNVs	*No. of SNVs in subset*	P
Coding nonsynonymous	11	5.7×10^{-4}
Conditioned on GWAS variants[a]		0.019
Coding nonsynonymous novel[b]	6	0.38
Noncoding predicted functional[c]	188	0.38
Noncoding predicted functional novel[b,c]	156	0.16

[a]Conditioned on proxies of rs2114912:G>T and rs35767:A>G.
[b]Not known in dbSNP, 1000 genomes project, or ESP 6500.
[c]Predicted transcription factor binding site (ENCODE ChipSeq, HMR) or DNAse hypersensitive site (ENCODE DHS) or ORegAnno regulatory variant or highly conserved (PhastCons).

Analyses of targeted sequence data

Table 3 and Supplementary Table 2 show descriptions of known and novel variants identified by targeted sequencing of the *IGF1* locus. Deep (mean read depth 50×) sequencing across the locus identified 1393 variants, 1143 (82.1%) of which were rare and novel. A total of 11 coding nonsynonymous variants were present, including 6 that were novel. Of the 1376 noncoding variants, 188 (14%) were predicted to be functional, including 156 that were novel. The large majority of the variants at the *IGF1* locus had MAF <0.1% (Supplementary Figure 3). Of all variants present at the locus, 893 (64%) were only observed one time in our samples. Of the novel variants, 198 (17%) were present in at least two of the three cohorts.

Meta-analyzed SKAT results (Table 4) showed that the subset of 11 rare coding nonsynonymous variants was significantly associated with ln(FI) ($P=5.7\times10^{-4}$). One rare variant (rs151098426:C>T, MAF=0.1%) accounted for 92.16% of the overall SKAT Q-statistic (Supplementary Table 3 and Supplementary Figure 4). This variant resulted in an alanine-to-threonine substitution and was predicted to be damaging by PolyPhen-2,[25] LRT,[26] and MutationTaster.[27] In contrast to the positive effect estimate for the rare T allele of rs151098426:C>T in the SKAT targeted sequencing analysis (Supplementary Table 3), 3 of the 1745 FHS offspring participants and 11 of the 3372 FHS generation 3 participants with follow-up genotyping of rs151098426:C>T carrying the rare allele had lower FI levels than the noncarriers (offspring: $\beta=-0.05$; generation 3: $\beta=-0.15$). These differences between carriers and noncarriers were nonsignificant (offspring: $P=0.734$; generation 3: $P=0.313$). The geometric means and the corresponding confidence intervals in carriers and noncarriers are shown in Supplementary Figure 5. Lookup of the variant in CHARGE exome chip results revealed a positive, but also nonsignificant, effect of rs151098426:C>T on FI levels (MAF= 0.14%, $\beta=0.02$, $P=0.471$).

Conditioning on proxies of the known FI GWAS variants rs2114912 and rs35767 attenuated the significant SKAT result to a nominal significant *P*-value ($P_{\text{conditioned}}=0.019$, Table 4), suggesting that the

GWAS signal explains part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of ENCODE Consortium regulatory element data sets and public transcriptome data in the UCSC Genome Browser suggested that GWAS variants in the vicinity of *IGF1* might have a direct functional role. In particular, rs35767 is ~1.2 kb upstream of the *IGF1* promoter and merely a few bases away from a strong FOXA1 binding site that is observed in ENCODE ChIP-seq data across a variety of human cell lines. Similarly, rs2114912:G>T is ~1.7 kb away from a strong ENCODE DNAseI hypersensitive site seen in multiple cell lines, including pancreatic islets, that overlaps an ENCODE transcription factor binding site ChIP-seq cluster for several transcription factors, including FOXA1. This combination of open chromatin as delineated by the DNAse I hypersensitive site with transcription factor binding in ChIP-seq data constitutes a regulatory element signature that warrants experimental validation. Rs2607988:G>A, a SNP in high LD with rs2114912:G>T ($r^2 = 0.882$ in HapMap2 CEU), is located in a ChIP-seq site for FOXA1 and alters a motif for FOXA. Interrogating the GTEx database, we did not find evidence for the GWAS variants to influence gene expression in any of the available tissues.

Single-variant analyses did not reveal significant associations with FI for any of the common variants present in the targeted sequence data (Supplementary Figure 6), including the proxies of the known FI GWAS hits rs35767:A>G ($P_{meta} = 0.69$) and rs2114912:G>T ($P_{meta} = 0.54$) (Supplementary Table 4), most likely because of the much smaller sample size in these targeted sequence data compared with the original, very large discovery sample sizes.

DISCUSSION

This study suggests that previously observed associations between SNVs near *IGF1* with FI levels were not mediated by circulating IGF-I levels. Further investigation of the *IGF1* gene, using deep sequencing data, revealed a large number of rare variants at the locus that had not been previously described, the large majority of which was very rare. A subset of rare coding nonsynonymous variants, including six novel variants and five variants that had been previously identified, was significantly associated with FI levels. Conditional analysis suggested that the common noncoding variants near *IGF1* that were identified in GWAS[4,5] explain part of the rare variant signal and the presence of a residual independent rare variant effect. Examination of ENCODE Consortium regulatory element catalogs showed that the GWAS variants were located in the proximity of FOXA1 binding sites and DNAseI hypersensitive sites, suggesting that they might have a direct functional role. This finding is noteworthy because FOXA1 is a key transcriptional regulator implicated in glucose metabolism and insulin secretion.[28,29] Studies in human cell culture and animal models will be needed to interrogate and validate the function of these noncoding variants in insulin biology.

One variant, rs151098426:C>T, resulting in an alanine-to-threonine substitution and predicted to be damaging by several annotation tools, seemed to drive the rare variant association. However, follow-up genotyping of rs151098426:C>T in an independent set of samples and lookup of the variant in CHARGE exome chip results did not reveal significant differences in FI levels between carriers and noncarriers of the rare allele, suggesting the absence of a single-variant effect for rs151098426:C>T on FI levels. Several recently published studies have demonstrated the need for large sample sizes to robustly identify associations of low-frequency variants with complex traits.[30–36] Because of the low MAF of rs151098426:C>T and thus the relatively small number of carriers, analyzing the variant in large numbers of additional samples will be required to definitively conclude whether this variant is associated with FI levels. Taking the FHS log FI distribution as an example and using a replication α of 0.05, if the effect was as large as we find in the SKAT results (1.32 SD), we would need 1657 samples to demonstrate the effect. However, this effect is likely to be an overestimate because of the winner's curse. If the effect was modest as we found in the FHS offspring (0.17 SD), a sample size of 97 128 would be needed.

We did not find a mediation effect of circulating IGF-I levels on the association of SNVs near *IGF1* with FI levels. However, measurement errors in IGF-I levels might be responsible for the absent observation of a mediation effect. Circulating IGF-I levels measured with an imperfect assay and at a single point in time may not sufficiently characterize the biologically relevant levels. However, although circulating levels of IGF-I decline with aging,[37] the levels do not undergo large short-term fluctuations.[38] Furthermore, in 3977 FHS participants, circulating IGF-I levels correlated negatively with insulin resistance, diabetes, and metabolic syndrome,[3] suggesting that these measures do represent biologically relevant levels and thus making measurement errors a less likely cause for not observing a mediation effect of IGF-I in our study.

The identification of variants at the *IGF1* locus that had not been previously described has increased our insight into the variation present at the locus. In line with previous sequencing studies,[34,39,40] we identified a large number of very rare variants, the majority (64%) even observed only one time in our samples. The presence of large numbers of very rare variants in the human genome is likely explained by recent explosive human population growth.[40,41] It has been hypothesized that these variants might harbor larger effects than those observed for common variants, as selection can have influenced only the most deleterious variants.[40] However, even for rare variants with larger effects, large sample sizes are needed to definitely conclude whether they influence complex traits because of the low MAF.

The strengths of this study in the CHARGE Targeted Sequencing framework include the high average sequence depth combined with stringent QC applied across the three cohorts, increasing confidence that even the rarest observed variation is real variation and not a technical artifact. Furthermore, we genotyped variant rs151098426:C>T in non-overlapping samples serving as replication cohort and as further evidence that the variant is real. A limitation of this study is type 2 error, both in mediation and targeted sequence analyses, where limited sample sizes have limited power to detect common and rare variant associations. The targeted sequence samples included only seven heterozygous carriers of the variant of interest rs151098426:C>T. With 3539 samples in this discovery set and a significance level of 0.001, for modest differences such as 0.1 SD in log FI, our power was 1% for MAF = 1% and 22% for MAF = 10%. Furthermore, because of the limited number of individuals with both targeted sequence data and IGF-I levels available in our study, it was not possible to test whether association of the subset of rare nonsynonymous variants with FI was mediated by IGF-I levels. Mean BMI was in the overweight range in all cohorts. However, evidence exists that effect sizes of known glycemic trait-associated variants do not differ between BMI strata.[5] As previously observed, FI values varied widely across studies, likely because of limited standardization across assays. Previous gene discovery studies, however, despite the same observation were successful in identifying FI-associated variants.[4,5] Finally, our study only included individuals of European ancestry, and this might

limit the generalizability to other ancestries of the observed *IGF1* variants and variant associations in this study.

In conclusion, our analyses suggest that association of SNVs near the *IGF1* gene with FI is not mediated by circulating IGF-I levels. Furthermore, our study increased insight into variation present at the *IGF1* locus and thus into the specific local coding as well as noncoding genetic architecture underlying FI levels, showing a large number of novel rare variants present at the locus and suggesting association of both rare coding nonsynonymous variants and a potential direct functional effect of common noncoding GWAS SNVs in the region on FI levels.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Funding support for 'Building on GWAS for NHLBI-diseases: the U.S. CHARGE Consortium' was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). Data for 'Building on GWAS for NHLBI-diseases: the U.S. CHARGE Consortium' were provided by Eric Boerwinkle on behalf of the Atherosclerosis Risk in Communities (ARIC) Study, L Adrienne Cupples, principal investigator for the Framingham Heart Study, and Bruce Psaty, principal investigator for the Cardiovascular Health Study. Sequencing was carried out at the Baylor Genome Center (U54 HG003273). The *ARIC Study* is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The *Framingham Heart Study* is conducted and supported by the NHLBI in collaboration with Boston University (Contract No. N01-HC-25195), and its contract with Affymetrix, Inc., for genome-wide genotyping services (Contract No. N02-HL-6-4278), for quality control by Framingham Heart Study investigators using genotypes in the SNP Health Association Resource (SHARe) project. A portion of this research was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resources at Boston University Medical Campus. Also supported by R01 DK078616 (Dr Meigs) and K24 DK080140 (Dr Meigs). This *CHS* research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, and N01HC85086; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through 1R01AG031890 and R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI Grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) Grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The generation and management of GWAS genotype data for the *Rotterdam Study* is supported by the Netherlands Organization for Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) project nr. 050-060-810, CHANCES (nr 242244). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Jones JI, Clemmons DR: Insulin-like growth factors and their binding proteins: biological actions. *Endocr Rev* 1995; **16**: 3–34.
2 Laughlin GA, Barrett-Connor E, Criqui MH, Kritz-Silverstein D: The prospective association of serum insulin-like growth factor I (IGF-I) and IGF-binding protein-1 levels with all cause and cardiovascular disease mortality in older adults: the Rancho Bernardo Study. *J Clin Endocrinol Metab* 2004; **89**: 114–120.
3 Lam CS, Chen MH, Lacey SM *et al*: Circulating insulin-like growth factor-1 and its binding protein-3: metabolic and genetic correlates in the community. *Arterioscler Thromb Vasc Biol* 2010; **30**: 1479–1484.
4 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.
5 Manning AK, Hivert MF, Scott RA *et al*: A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 2012; **44**: 659–669.
6 Consortium EP, Bernstein BE, Birney E *et al*: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.
7 Karolchik D, Barber GP, Casper J *et al*: The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 2014; **42** (Database issue): D764–D770.
8 Psaty BM, O'Donnell CJ, Gudnason V *et al*: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* 2009; **2**: 73–80.
9 Lumley T, Dupuis J, Rice KM *et al*: Two-phase subsampling designs for genomic resequencing studies, 2012. Available from http://stattech.wordpress.fos.auckland.ac.nz/files/2012/05/design-paper.pdf.
10 Lin H, Wang M, Brody JA *et al*: Strategies to design and analyze targeted sequencing data: cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. *Circ Cardiovasc Genet* 2014; **7**: 335–343.
11 Servin B, Stephens M: Imputation-based analysis of association studies: candidate regions and quantitative traits. *PloS Genet* 2007; **3**: e114.
12 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
13 Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**: 186–194.
14 1000 Genomes Project Consortium, Abecasis GR, Altshuler D *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
15 Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. Available from http://evs.gs.washington.edu/EVS/ (Accessed via ANNOVAR).
16 Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164.
17 Siepel A, Bejerano G, Pedersen JS *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005; **15**: 1034–1050.
18 Griffith OL, Montgomery SB, Bernier B *et al*: ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* 2008; **36**(Database issue): D107–D113.
19 GTEx Consortium: The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45**: 580–585.
20 Lumley T: rmeta: Meta-analysis. R package version 2.16, 2012. Available from http://CRAN.R-project.org/package=rmeta.
21 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.
22 Chen H, Meigs JB, Dupuis J: Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 2013; **37**: 196–204.
23 Willer CJ, Li Y, Abecasis GR: METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–2191.
24 Li J, Ji L: Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 2005; **95**: 221–227.
25 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
26 Chun S, Fay JC: Identification of deleterious mutations within three human genomes. *Genome Res* 2009; **19**: 1553–1561.
27 Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–576.
28 Gao N, Le Lay J, Qin W *et al*: Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. *Mol Endocrinol* 2010; **24**: 1594–1604.
29 Kaestner KH: The FoxA factors in organogenesis and differentiation. *Curr Opin Genet Dev* 2010; **20**: 527–532.
30 Kozlitina J, Smagris E, Stender S *et al*: Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2014; **46**: 352–356.
31 Carty CL, Spencer KL, Setiawan VW *et al*: Replication of genetic loci for ages at menarche and menopause in the multi-ethnic Population Architecture using Genomics and Epidemiology (PAGE) study. *Hum Reprod* 2013; **28**: 1695–1706.
32 Peloso GM, Auer PL, Bis JC *et al*: Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* 2014; **94**: 223–232.

33 Huyghe JR, Jackson AU, Fogarty MP *et al*: Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 2013; **45**: 197–201.
34 Flannick J, Thorleifsson G, Beer NL *et al*: Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* 2014; **46**: 357–363.
35 Holmen OL, Zhang H, Fan Y *et al*: Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat Genet* 2014; **46**: 345–351.
36 Zuk O, Schaffner SF, Samocha K *et al*: Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci USA* 2014; **111**: E455–E464.
37 Kaplan RC, Buzkova P, Cappola AR *et al*: Decline in circulating insulin-like growth factors and mortality in older adults: cardiovascular health study all-stars study. *J Clin Endocrinol Metab* 2012; **97**: 1970–1976.
38 Ketha H, Singh RJ: Clinical assays for quantitation of insulin-like-growth-factor-1 (IGF1). *Methods* 2015; **81**: 93–98.
39 Morrison AC, Voorman A, Johnson AD *et al*: Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* 2013; **45**: 899–901.
40 Coventry A, Bull-Otterson LM, Liu X *et al*: Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 2010; **1**: 131.
41 Keinan A, Clark AG: Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012; **336**: 740–743.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)