Twin Research and Human Genetics
Volume 18 ▌ Number 2 ▌ pp. 117–125 ▌ © The Author(s) 2015 ▌ doi:10.1017/thg.2015.10

Exome Sequencing to Detect Rare Variants Associated With General Cognitive Ability: A Pilot Study

Michelle Luciano,[1,2] Victoria Svinti,[3] Archie Campbell,[4] Riccardo E. Marioni,[1,4,5] Caroline Hayward,[3] Alan F. Wright,[3] Martin S. Taylor,[3] David J. Porteous,[1,4] Pippa Thomson,[1,4] James G.D. Prendergast,[3] Nicholas D. Hastie,[3] Susan M. Farrington,[3] Generation Scotland,[6] Malcolm G. Dunlop,[3] and Ian J. Deary[1,2]

[1]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK
[2]Department of Psychology, University of Edinburgh, Edinburgh, UK
[3]MRC Human Genetics, Unit MRC IGMM, University of Edinburgh, Edinburgh, UK
[4]Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[5]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia
[6]A collaboration between the University Medical Schools and National Health Service in Aberdeen, Dundee, Edinburgh and Glasgow, UK

Variation in human cognitive ability is of consequence to a large number of health and social outcomes and is substantially heritable. Genetic linkage, genome-wide association, and copy number variant studies have investigated the contribution of genetic variation to individual differences in normal cognitive ability, but little research has considered the role of rare genetic variants. Exome sequencing studies have already met with success in discovering novel trait-gene associations for other complex traits. Here, we use exome sequencing to investigate the effects of rare variants on general cognitive ability. Unrelated Scottish individuals were selected for high scores on a general component of intelligence (*g*). The frequency of rare genetic variants (in $n = 146$) was compared with those from Scottish controls (total $n = 486$) who scored in the lower to middle range of the *g* distribution or on a proxy measure of *g*. Biological pathway analysis highlighted enrichment of the mitochondrial inner membrane component and apical part of cell gene ontology terms. Global burden analysis showed a greater total number of rare variants carried by high *g* cases versus controls, which is inconsistent with a mutation load hypothesis whereby mutations negatively affect *g*. The general finding of greater non-synonymous (vs. synonymous) variant effects is in line with evolutionary hypotheses for *g*. Given that this first sequencing study of high *g* was small, promising results were found, suggesting that the study of rare variants in larger samples would be worthwhile.

■ **Keywords:** generation Scotland: the Scottish family health study, intelligence, IQ, genetics

Determinants of cognitive function are important to understand because of the significant impact they might have on physical and mental health traits and social outcomes with which intelligence is positively associated (Carroll et al., 2011; Deary & Batty, 2007; Gale et al., 2008). The presumed survival and reproductive advantages of higher cognitive ability suggest it to be a trait under positive selection, at least until several centuries ago (Miller, 2000; Miller & Penke, 2007). The heritability of cognitive ability estimated from twin designs shows estimates of around 0.30 in childhood, which increase to as much as 0.80 in adulthood (Briley & Tucker-Drob, 2013; Haworth et al., 2010; McGue & Christensen, 2013). The present study focuses on the contribution of rare variants to this genetic variation.

Molecular genetic studies of cognitive ability started with genome-wide linkage analysis (Posthuma et al., 2005) and candidate gene studies that have culminated in very few —

RECEIVED 26 September 2014; ACCEPTED 5 February 2015. First published online 6 March 2015.

ADDRESS FOR CORRESPONDENCE: Michelle Luciano Psychology, University of Edinburgh, 7 George Square, EH8 9JZ, Scotland, UK. E-mail: michelle.luciano@ed.ac.uk

if any — replicated associations (Chabris et al., 2012). More recent genome-wide association scans have failed to detect any genome-wide significant single nucleotide polymorphism (SNP) associations (Benyamin et al., 2014; Davies et al., 2011). The largest investigations of structural genetic variants — more specifically, rare copy number variants (CNVs) — in samples of adolescents and older adults failed to show correlations between CNV number, length or number of genes disrupted with IQ (MacLeod et al., 2012; McRae et al., 2013). There has been only one study of rare single nucleotide genetic variants and cognitive ability (Marioni et al., 2014b). These classes of variants are more recent than common variants or have drifted to low frequency, so are of increased population specificity (Abecasis et al., 2012).

The literature to date points to common SNPs as the main type of additive genetic variation contributing to general cognitive ability (Davies et al., 2011; Marioni et al., 2014a; Trzaskowski et al., 2014). The genetic contribution is well described by a large number of variants contributing very small amounts (less than 0.5%) of individual variation. For example, Marioni et al. (2014a) used pairwise genetic relationships between 6,609 individuals (drawn from the same wider sample reported in our study) based on 594,824 autosomal SNPs to predict similarity in pairwise intelligence scores. They estimated that 29% (standard error of 0.05) of the variation in general intelligence could be explained by linkage disequilibrium between the genotyped SNPs and unknown causal markers. This leaves a substantial proportion of additive genetic variance unexplained; rare genetic variants are a potential source of this missing heritability. If, unlike common variants, rare genetic variants have large effects on intelligence variation, then they could provide tractable variants to study because they are more likely to have functional effects than common SNPs.

The only study to date on rare genetic variants — usually defined as those occurring with less than 0.5 or 1% frequency in a population — examined exome (protein coding regions) chip variants in relation to normal variation in cognitive ability measured in childhood and old age in a Scottish population (Marioni et al., 2014b). The presence of rare variants (<1% frequency) in an individual were summed into a global burden score and this was used to predict childhood and old age general cognitive ability (*g*), in line with the hypothesis that mutation load is detrimental to *g*. No significant associations were found, including separate tests of stop-gain/loss, splice and missense mutations. This study was limited in that (1) it focused on exome chip data, where variants have been identified via a reference sample and thus may miss highly relevant population-specific variants, and (2) it used population samples that were relatively high in average ability and restricted in cognitive variance. The present study overcomes these limitations by using sequencing methods and by focusing on a very high cognitive ability extreme group (>2.3 *SD* from the mean). Whereas the technology to sequence DNA has rapidly advanced, it is still relatively expensive to perform whole-genome sequencing. A cheaper alternative is to sequence exons of genes, which comprise between 1 and 2% of the human genome (Ng et al., 2009; Tennessen et al., 2012). Because there is stronger negative selection for rare alleles in coding versus intergenic regions, these regions are more likely to harbor functional variants and thus relate to an evolutionarily important trait-like cognitive ability.

Here, we investigate whether rare variants of moderate-to-large effect might influence cognitive ability using a dichotomized trait approach. We compare a group of individuals with a very high cognitive ability (cases) with 'controls' of low-to-average cognitive ability. Extremely low cognitive ability individuals were not sampled because they might capture syndromic intellectual disability; any detectable rare variant effects would then be inseparable from other neurological/physical comorbidities and not generalizable to normal varying *g*. We conduct variant-by-variant and gene-based analyses.

Materials and Methods

Samples

Generation Scotland: The Scottish Family Health Study (GS: SFHS). This is an extended pedigree cohort of families mainly residing in the Glasgow, Tayside, and Grampian regions of Scotland. Details of recruitment and testing of this population-based cohort, and of some distributions and associations of the variables tested, can be found in Smith and colleagues (Smith et al., 2006; 2013). Ethical approval for this cohort study was granted by the NHS Tayside Committee on Medical Research Ethics Committee (REC Reference Number: 05/S1401/89) and Research Tissue Bank status by the Tayside Committee on Medical Research Ethics Committee (REC Reference Number: 10/S1402/20).

Individuals were ascertained through participating general medical practitioners. They were not selected for medical disease/disorder status. These probands then recruited their family members. More than 21,500 individuals were measured on cognitive, personality, and physical/mental health variables between 2006 and 2011 (2013; 2006).

Of relevance to this study were their scores on the cognitive tests. A principal components analysis of Logical Memory (summed immediate and delayed scores from one paragraph) from the Wechsler Memory Scale III (Wechsler, 1998), Digit Symbol-Coding from the Wechsler Adult Intelligence Scale III (Wechsler, 1997), phonemic Verbal Fluency (letters C, F, L; Lezak, 2004) and Mill Hill Vocabulary (combined junior and senior synonyms; Raven et al., 1977) produced a first unrotated principal component that explained 42% of the total variance. Component loadings (weighted most heavily on the verbal fluency and vocabulary tests) were used to construct a composite general cognitive ability (*g*) score for each individual that was then regressed on age. The component score was used for selection because

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

we were interested in general cognitive processes, but it has the added advantage of reducing noise stemming from test measurement error. The age-residualized component scores were ranked and the top 76 female scores and top 74 male scores were selected to be high *g* cases for sequencing. After sequencing dropout, 146 (76 female) individuals remained. These scores ranged 2.34–3.97 standard deviations above the sample mean for *g* (mean 2.76, *SD*: 0.36).

Three hundred and thirty-one controls were also selected from GS: SFHS. They were a readily available sample that came from concurrent sequencing studies of two other selected traits. They included 81 (56 female) individuals with depression and 27 (11 female) non-affected relatives of individuals with depression, and 223 (160 female) mostly obese individuals (∼ 95% with BMI >40 plus parents with very low BMI). These controls were selected from a larger sequenced sample of 502 based on their age- and sex-residualized standard *g* scores being at least 2 standard deviations below the minimum score of the high *g* group. Their scores ranged from -4.11 to 0.33 (mean -0.72, *SD*: 0.79); only 13 individuals had *g* scores between -2.5 and -4.11, so it was unlikely that this sample was enriched for intellectual impairment which, as mentioned, might have a different genetic etiology to normal cognitive ability. All the included participants were unrelated: relatedness was based on reported relationships between cohort members. For related persons, the individual with the lower *g* score was selected, and obese individuals were chosen over depressed individuals within the same pedigree because they come from a larger (more representative) population. The mean age of cases was 49.8 years (*SD*: 11; range: 22–69) and of controls was 47.7 years (*SD*: 13.9; range: 18–80). Of the high *g* cases, 20 had (Structural Clinical Interview) diagnoses of major depressive or bipolar disorder, and of the controls, 146 had these diagnoses. The range of BMI in cases was 18.41 to 49.18 (mean 26.11 ± 4.63) and in controls it was 15.25 to 67.18 (mean 38.49 ± 9.57).

Colorectal Cancer (CRC) Study. A further 155 controls were drawn from ongoing studies into susceptibility to CRC in the Scottish population (Barnetson et al., 2006; Liu et al., 1995). The patients were selected based on having a 1st degree relative affected with CRC (98%), and largely excluded participants with known gene mutations in *APC*, *MUTYH* and Mismatch repair defects where possible. This sample set did not have cognitive test scores, so a proxy measure of cognitive ability based on their attained education level and/or Carstairs and Morris index of deprivation (http://www.nhslothian.scot.nhs.uk/publichealth/2005/ar2003/dataset/depcats.html) was used to select probable lower *g* participants. The samples selected to use as controls comprised of 123 (58 female) individuals who did not report any education in the form of senior (or advanced) high school certificate or equivalent, diplomas, degrees or professional qualifications, and 32 (16 female) reporting a high school standard grade senior certificate equivalent with high/intermediate (socio-economic) deprivation. Attained education (in particular) and SES are both established correlates of cognitive ability (Luciano et al., 2010; Strenze, 2007). The age range of the selected sample was 27 to 79 years, with a mean of 64.8 ± 8.96 years.

DNA Collection

In GS: SFHS, DNA was primarily obtained by blood sampling with 35 saliva samples collected in the clinic. Samples were processed by the Wellcome Trust Clinical Research Facility Genetics Core, Edinburgh (Kerr et al., 2013). Whole blood DNA was extracted in the CRC study using NucleonTM in house or by the WTCRF Genetics Core.

Sequencing Method

The GS: SFHS high *g* cases and depression controls were processed at the Edinburgh Genomics facility at the University of Edinburgh. Exon capture was carried out using the Illumina TruSeq exome enrichment kit and resulting libraries sequenced on an Illumina HiSeq machine generating 100bp paired-end reads. The mean coverage per sample was approximately 38×. GS: SFHS obesity controls were sequenced at the Wellcome Trust Sanger Institute using the Agilent SureSelect kit having an average coverage of 86× per sample. The CRC samples were sequenced by Edinburgh Genomics using a similar protocol to GS: SFHS high *g* cases, with a mean coverage per sample of 39×.

Sequenced reads were aligned to the hg19 (b37) 1,000 genomes reference using BWA 0.5.9 (Li & Durbin, 2009). Duplicate reads were marked using Picard 1.79. Samtools 0.1.16 (Li et al., 2009) was used at various steps along this pipeline. Realignment around indels, base quality score recalibration and variant discovery were performed using the Genome Analysis Tool Kit (GATK) version 2.7.2, according to GATK Best Practices recommendations for exome sequence analysis (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). SNP and INDEL discovery were carried out using GATK's UnifiedGenotyper across all samples simultaneously, using reduced reads and downsampling. The search was restricted to regions covered by the exome capture kits with 50bp flanking sequence. Variant recalibration was carried out with GATK 2.8.1 due to improved performance on our dataset, and snpEff 3.3 was used for annotating variants (Cingolani et al., 2012). Given that different exome capture platforms were used in these studies (Illumina TruSeq and Agilent SureSelect), only regions covered by both platforms were considered for downstream analysis in order to minimize technical artifacts. Quality control was carried out using the PLINK software (Purcell et al., 2007) with the following criteria: genotype rate per site ≥ 95%, genotyping rate per sample ≥ 90%, and excluding samples with extreme heterozygosity (± 3 *SD* from the mean).

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

Statistical Analysis

Case versus control single variant analysis was performed primarily to check whether gene-based results were not being unduly influenced by a single variant association. The small sample size limited the statistical power of these tests, with a Bonferroni correction (for multiple variant testing of polymorphic SNPs < 5%) giving the following corrected genome-wide significance levels: 1.63×10^{-7} (combined controls), 2.78×10^{-7} (depression controls), 2.25×10^{-7} (obesity controls), and 2.60×10^{-7} (cancer controls). These analyses were performed in PLINK (Purcell et al., 2007) using Fisher's exact test.

Gene-based case-control analyses were performed using the optimized sequence kernel association test (SKAT-O; Lee et al., 2012). SKAT-O is a gene-based test that optimally combines the classic burden test with the non-burden SKAT test in order to maximize power. Such a test then allows a mixture of hypotheses, that is, that most rare variants in a gene are causal and unidirectional or that most variants in a gene are non-causal and have varying directions of effect. The test also implements a small sample adjustment method that was useful for the current study, given the sizes of our sample sets. The main analysis combined all sample controls in an attempt to remove any systematic bias caused by controls having been selected for depression, obesity or cancer. Subsequent analyses were also performed using depressed, obese or cancer samples as separate control groups to check for consistency in results across different sampling frames. Only results that were consistent across all three samples were taken as evidence of an effect on cognitive ability; any significant result in a single analysis could reflect variants for either cognitive ability or the trait sampled for in the controls. And in the obesity and cancer control analyses, any unique results could also be due to sequencing batch effects because the sequencing of cases and controls was not performed together, thus consistency in signal across all three analyses was sought. Separate analyses were carried out using different inclusion criteria for single nucleotide variants. Filters were based on allele frequencies of (a) less than 1% and (b) less than 5%, including (1) all types of variants, (2) non-synonymous/splice/frameshift mutations, and (3) synonymous coding variants. We test both non-synonymous and synonymous coding variants because they have been shown to have a comparable likelihood of association and effect size based on a comparative analysis of 2,113 published genome-wide association studies (Chen et al., 2010). For each SKAT-O analysis, 1,000 permutations were performed and these were used to estimate the Family Wise Error Rate (FWER). An FWER threshold of 0.05 was used to determine significance.

Results from the gene-based analyses were ranked in ascending order by their *p*-value and processed using GOrilla (Eden et al., 2009), a web-based pathways analysis program that uses the minimum hypergeometric score to identify gene ontology enrichment for genes at the top of a ranked list. This ranking of all genes circumvents the need to use arbitrary significance level cut-offs to define a target set of genes that is enriched for GO terms compared to a background set of genes. *P*-values less than 10^{-3} were requested and corrected for multiple testing, producing a false discovery rate q-value, with values < 0.05 deemed significant. Any gene sets containing a single variant were excluded from this analysis.

A final analysis testing the difference in genome-wide burden (i.e., total number of minor alleles carried by an individual) of rare/low frequency variants (total and non-synonymous with < 1% and < 5% frequency) between cases and combined controls was performed in R (R Core Team, 2014) using an independent groups *t*-test.

Detection sensitivity between cohorts was calculated using the estimations from Meynert and colleagues (2013). For each sample, the depth of coverage for the analyzed regions was obtained using GATK. The cumulative coverage counts (reported at depths 0–500), together with the genomic target size, were used to obtain the proportion of sites at each depth. The results of the product between estimated values for heterozygous sites from Meynert et al., (2013) and the proportion of sites, at each depth, were summed to get an overall SNP detection sensitivity per sample. The average of detection sensitivities per cohort was obtained by averaging the individual values across samples in each cohort.

Despite our sample being homogenous in terms of their Scottish ancestry, population stratification was assessed by deriving four multidimensional scaling (MDS) components in PLINK (Purcell et al., 2007); separate derivations were performed for rare (allele frequencies < 5%) and common (> 5% MAF) variants. None of the MDS components for rare or common variants were associated with affection status in the combined sample ($p > .05$) nor were they associated with sequencing batch ($p > .05$), and thus were not used in further analyses.

Results

Single Variant Test

A significant, single variant result was found in the subanalyses of depression and obesity controls for rs4449373 (on chromosome 4), with respective *p*-values of 3.36×10^{-8} and 6.43×10^{-12}. The minor allele (T) was more frequent in cases (0.103) than controls (depression: 0; obesity: 0.002). The lowest *p*-value for the combined controls was 4.48×10^{-6} for rs200302560 (chromosome 14), and for the cancer controls was 2.35×10^{-5} for rs116314157 (chromosome 2).

Gene-Based Test of All Variants

The analysis of all single nucleotide variants with less than 1% frequency in the cases versus combined controls included 24,514 genes sets comprising 339,231 variants. No significant associations were found after FWER correction. A pathways analysis revealed no enrichment for gene

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

TABLE 1

Significant Gene Ontology Pathways Enriched in the Varying Analyses Comprising the Combined Controls

	Gene ontology	*p*-value	FDR *p*-value	Enrichment values*	Genes in pathway
Non-synonymous					
SNVs < 0.01/ < 0.05					
Cellular component GO:0031305	Integral component of mitochondrial inner membrane	4.6E-6/ 3.6E-6	.006/ .005	24.67 [13, 519, 10, 274, 5]/ 25.95 [13, 803, 10, 266, 5]	*TIMM23* — translocase of inner mitochondrial membrane 23 homolog (yeast) *COX18* — cox18 cytochrome c oxidase assembly factor *MCU* — mitochondrial calcium uniporter *ETFDH* — electron-transferring-flavoprotein dehydrogenase *TMEM11* — transmembrane protein 11
SNVs < 0.05					
Cellular component GO:0045177	Apical part of cell	6.1E-5	.04	5.99 [13, 803, 67, 344, 10]	*SLC11A2* — solute carrier family 11 (proton-coupled divalent metal ion transporter), member 2 *BTD* — biotinidase *EPB41L4B* — erythrocyte membrane protein band 4.1 like 4b *MYL12B* — myosin, light chain 12b, regulatory *VAMP7* — vesicle-associated membrane protein 7 *ITGA8* — integrin, alpha 8 *MYO6* — myosin vi *SLC25A27* — solute carrier family 25, member 27 *DVL2* — dishevelled segment polarity protein 2 *PARD6B* — par-6 partitioning defective 6 homolog beta (*C. elegans*)

Note: *p*-value, FDR corrected *p*-value, enrichment values, and prominent genes in each pathway are listed.
SNV: single nucleotide variants; *Enrichment is defined as (b/n)/(B/N). *N*: total number of genes; B: total number of genes associated with a specific GO term; *n:* number of genes in the 'target set'; b: number of genes in the 'target set' associated with a specific GO term.

ontology terms (16,205 genes associated with a GO term) after FWER correction. Subanalyses of the separate control cohorts did not find consistent effects, although three gene sets withstood FWER correction for depression controls (see Supplementary Table 1 — available on the Cambridge Journals Online website). Biological pathways analyses showed no significant results for the separate gene-based results of high *g* cases versus obesity, depression, and cancer controls.

No genes showed significant association when including variants with < 5% frequency (380,362) in the (24,699) gene sets. In the subanalyses of obesity, depression and cancer controls, gene sets were significant for obesity (three genes) and depression (five genes) controls, with three genes in the same locus overlapping (RP11-673E1.4, *GYPB*, *GYPA*; see Supplementary Table 2) between analyses. For the 15 variants at this locus, the rare allele was less frequent in cases for 8 variants. In the combined sample, this locus was nominally significant ($p \leq .005$). Other significant gene loci did not overlap between analyses and mostly contained either one or two variants in the gene set, making these results unlikely to represent true effects (i.e., none of them withstood correction at a single variant level). Genomic inflation (at both allele frequencies) was indicated in the tests of depression (lambda value ~ 1.28) and obesity (~ 1.31) controls but not in the cancer (~ 1.05) and combined (~ 1.07) controls. All biological pathways tests (combined and subsamples) were non-significant.

Gene-Based Test of Non-Synonymous, Splice and Frameshift Variants

When considering variants with a frequency less than 1% (134,751 variants in 20,791 gene sets), no genes reached FWER significance in the analysis of combined controls or cancer controls. For the obesity and depression controls analyses, one and eight genes, respectively, reached FWER significance, and *SYNGAP1* was significant in both analyses (see Supplementary Table 1). *SYNGAP1* showed a lack of variants in cases compared to controls from the combined sample, but the *p*-value was greater than 0.05. These variants were not present in cancer controls. When using the combined controls in a gene ontology analysis, integral component of mitochondrial inner membrane showed significant (FDR $p = .006$) enrichment of the 13,519 genes associated with a GO term. These results, including the prominent genes in each pathway, are shown in Table 1.

Results for variants with a frequency less than 5% (147,111 variants in 21,022 gene sets) overlapped somewhat with those less than 1% frequency for depression but not obesity analyses (see Supplementary Table 2). The locus RP11-673E1.4/*GYPB*/*GYPA* overlapped between obesity and depression control analyses, and this locus was nominally significant in the combined sample ($p < .001$). Four of the six variants were more frequent in cases than controls. Genomic inflation (at both allele frequencies) was again observed in the tests of depression (lambda value

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

~ 1.33) and obesity (~ 1.33) controls but not in the cancer (~ 0.99) and combined (~ 1.1) controls. Among the 13,803 genes associated with a GO term in the combined controls, integral component of mitochondrial inner membrane was again significantly enriched (FDR $p = .005$) and so too was the apical part of cell component (FDR $p = .04$; see Table 1). No pathways were enriched for the gene results in the subanalyses.

Gene-Based Test of Synonymous Variants

No gene associations were found in the combined, cancer or obesity controls analyses using variants with a frequency less than 1% (73,738 variants in 18,533 gene sets) or less than 5% (84,374 variants in 19,135 gene sets). For variants with a frequency less than 1%, five genes withstood FWER correction for the depression controls analyses (see Supplementary Table 1). For variants with a frequency less than a 5%, four genes withstood FWER correction for depression controls analyses (see Supplementary Table 2); and three of these overlapped with the 1% frequency results. Genomic inflation (at both allele frequencies) was observed in the tests of depression (lambda value ~ 1.27) and obesity (~ 1.26) controls but not in the cancer (~ 0.96) and combined (~ 1.12) controls. One gene ontology term, acetylgalactosaminyltransferase activity, was significantly enriched ($p = .03$) for the depression controls at the 1% allele frequency threshold (see Supplementary Table 3).

Genome-Wide Burden

The range of total minor alleles with less than 1% frequency per individual was between 765 and 2,544. The mean number of total minor alleles carried by high *g* cases ($M = 953.11$, $SD = 102.74$) was higher than controls ($M = 933.11$, $SD = 87.56$); Welch's $t\,(212) = 2.13$, $p = .03$. The range of total minor alleles with less than 5% frequency per individual was 2,265 and 4,479. There was a significant difference in the mean number of total minor alleles carried by high *g* cases ($M = 2{,}564.18$, $SD = 126.76$) and controls ($M = 2{,}537.56$, $SD = 123.56$); Welch's $t\,(234) = 2.24$, $p = .03$. A burden test including only non-synonymous variants (total number ranging 614 to 1,192 for < 5% frequency and 175 to 705 for < 1% frequency) was not significant ($p > .05$).

The average heterozygous detection sensitivity in analyzed regions is estimated (Meynert et al., 2013) to be almost identical in high *g* (95.5%) and control cohorts (95.9%). Consequently, the excess rare variants consistently found in the high *g* cases cannot be explained by acquisition bias through sequence depth or uniformity differences between cohorts.

Discussion

In this study, we have investigated the effect of rare/low frequency genetic variant differences between groups of people with very high general cognitive ability (*g*) and low-to-average *g*. No significant genes were supported. The integral component of mitochondrial inner membrane component was identified in the gene ontology enrichment analysis of combined control groups for variants at < 1% and < 5% allele frequencies. A more powerful genome-wide burden analysis showed that high *g* cases carried more rare variants (all types) than controls at both allele frequencies.

The study was limited in power due to the small sample size, and by the lack of a random control sample. By combining each of the selected control groups we wished to minimize systematic confounding of possible rare variant contributions to obesity, depression or cancer. The reduction of the genomic inflation factor when the control groups were combined suggests that this was effective; and that the significant results for the depression and obesity control groups might be influenced by confounding. Although we cannot fully rule out that sequencing batch effects influenced the results in the combined analysis, our analysis of MDS components did not identify any evidence of structure within the dataset linked to batch or exome enrichment kit. The combined analysis did not reveal any significant genetic loci for high *g* for any type of variant at either less than 1% or 5% allele frequency. However, the results of non-synonymous variants (at both 1% and 5% frequency) converged on a significant biological pathway, the integral component of mitochondrial inner membrane. The mitochondrial inner membrane contains many proteins with functions in energy conservation and protein and metabolite transport (Becker et al., 2009). Of the genes highlighted in this pathway, lower TIMM23 expression in knockout mice has been associated with poorer neurological functioning and reduced lifespan (Ahting et al., 2009).

For variants with a frequency less than 5%, the apical part of cell pathway was also significant: in the cytoskeleton, apical surfaces of epithelial cells have increased surface area affecting the absorption rate of nutrients (Costanzo, 2009). Of the significant genes in this pathway, Slc11a2 deficiency has been associated with hippocampal iron levels and related cognitive and neurodevelopmental traits in rodents (Carlson et al., 2010; Pisansky et al., 2013). *VAMP7* is suggested to play a role in neural transmission (Bal et al., 2013). SNPs in *ITGA8* have been associated with schizophrenia risk, a disorder characterized by cognitive dysfunction (Supriyanto et al., 2013). Variants in *SLC25A27* have been associated with schizophrenia (Mouaffak et al., 2011; Yasuno et al., 2007) and with *SLC25A27* expression changes observed in schizophrenia post-mortem brain (Chu & Liu, 2010). Reduced *SLC25A27* expression has been found in the anterior cingulate gyrus, motor cortex, and thalamus in autism post-mortem brain (Anitha et al., 2012). In mouse neocortex, blocked Dvl2 expression decreases proliferative and neurogenic neural progenitor cells (Endo et al., 2012). And a SNP in *PARD6B* has been associated with bipolar disorder in a very small GWAS in the Bulgarian population (Yosifova

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

et al., 2011). Our results for synonymous variants in the combined analysis did not highlight any biological pathway, suggesting that non-synonymous variants play a more important role in influencing cognitive ability, as would be predicted by evolutionary theory.

In another attempt to overcome any systematic confounding due to control group selection, we performed separate control group analyses and checked for consistency between results. Three overlapping gene loci in the same locus — RP11-673E1.4, *GYPB* (glycophorin B; MNS blood group), *GYPA* (glycophorin A; MNS blood group)) — were significant for the depressed and obesity control subanalyses. These were the two groups selected for low-to-average *g* using the same cognitive tests as in high *g* group selection, so they might represent better control groups than the cancer controls (where this gene locus was not detected). But because consistency was not observed across all three cohorts, we will not speculate further on the involvement of these genes.

The results of the genome-wide burden analyses were significant for all types of variants, which was inconsistent with Marioni et al.'s (2014b) null findings of rare/infrequent exome chip variant burden and *g* in childhood or old age. Furthermore, our finding was that increased rare variant burden related to higher *g*, which is inconsistent with a mutational-selection balance (i.e., the balance between new harmful mutations arising and selection against them) explanation for *g* (Penke et al., 2007). However, non-synonymous variants are arguably more relevant to the mutational load hypothesis and we did not find a difference between cases and controls for this type of rare variant burden. Because our comparison was between high *g* and low to normal cognitive ability, we cannot dismiss the role of non-synonymous rare variant burden in influencing *g* at the very low extreme. Without replication, it may be that our finding of higher *g* being associated with increased burden of all types of rare variants represents type 1 error, especially given that we observed greater variance in rare variant burden in the smaller high *g* group.

The present pilot study suggests that further exploration of the contribution of rare variants to *g* should be pursued. Larger samples, such as the one currently being sequenced by BGI Cognitive Genomics Lab (focusing on 2,236 cases with extremely high *g*, i.e., > 150 IQ points) will be well-powered to investigate rare variants at a gene based, and perhaps even single-variant level (L.C.A.M. Tellier, personal communication). Additionally, their study involves whole genome sequencing rather than restricted to the exome, as here. Because some regions of non-coding DNA are highly conserved (~ 3% among distantly related mammals), and have been shown to be under purifying selection in humans, they are potentially rich in functional variants (Drake et al., 2006), and particularly relevant to a trait such as *g*. The sequencing of families will be required to measure the effect of de novo mutations, which are not identifiable in population based studies like ours; such effects might contribute to the substantial non-shared environmental variance in *g*.

Acknowledgments

We thank the GS: SFHS and CRC participants, and all people involved in phenotypic data collection and biological sample processing. We are grateful to Veronique Vitart for her contribution to the quality control of GS: SFHS sequencing data. The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (MR/K026992/1). Funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council (MRC) is gratefully acknowledged. GS: SFHS was funded by a grant from the Scottish Government Health Department, Chief Scientist Office, number CZD/16/6. The MRC Human Genetics Unit QTL Group funded GS: SFHS high *g* and depression exome sequencing. The obesity control sequencing makes use of data generated by the UK10K Consortium; a full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. Programme Grant funding from Cancer Research UK (C348/A12076) funded cancer exome sequencing.

Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/thg.2015.10.

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., ... McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*, 56–65.

Ahting, U., Floss, T., Uez, N., Schneider-Lohmar, I., Becker, L., Kling, E., ... Klopstock, T. (2009). Neurological phenotype and reduced lifespan in heterozygous Tim23 knockout mice, the first mouse model of defective mitochondrial import. *Biochimica et Biophysica Acta, 1787*, 371–376.

Anitha, A., Nakamura, K., Thanseem, I., Yamada, K., Iwayama, Y., Toyota, T., ... Mori, N. (2012). Brain region-specific altered expression and association of mitochondria-related genes in autism. *Molecular Autism, 3*, 12.

Bal, M., Leitz, J., Reese, A. L., Ramirez, D. M. O., Durakoglugil, M., Herz, J., ... Monteggia, L. M. (2013). Reelin mobilizes a VAMP7-dependent synaptic vesicle pool and selectively augments spontaneous neurotransmission. *Neuron, 80*, 934–946.

Barnetson, R. A., Tenesa, A., Farrington, S. M., Nicholl, I. D., Cetnarskyj, R., Porteous, M. E., ... Campbell, H. (2006). Identification and survival of carriers of mutations in DNA

mismatch-repair genes in colon cancer. *New England Journal of Medicine, 354*, 2751–2763.

Becker, T., Gebert, M., Pfanner, N., & van der Laan, M. (2009). Biogenesis of mitochondrial membrane proteins. *Current Opinion in Cell Biology, 21*, 484–493.

Benyamin, B., Pourcain, B., Davis, O. S., Davies, G., Hansell, N. K., Brion, M. J., ... Visscher, P. M. (2014). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry, 19*, 253–258.

Briley, D. A., & Tucker-Drob, E. M. (2013). Explaining the increasing heritability of cognitive ability across development: A meta-analysis of longitudinal twin and adoption studies. *Psychological Science, 24*, 1704–1713.

Carlson, E. S., Fretham, S. J., Unger, E., O'Connor, M., Petryk, A., Schallert, ... Georgieff, M. K. (2010). Hippocampus specific iron deficiency alters competition and cooperation between developing memory systems. *Journal of Neurodevelopmental Disorders, 2*, 133–143.

Carroll, D., Batty, G. D., Mortensen, L. H., Deary, I. J., & Phillips, A. C. (2011). Low cognitive ability in early adulthood is associated with reduced lung function in middle age: The Vietnam experience study. *Thorax, 66*, 884–888.

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., ... & Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science, 23*, 1314–1323.

Chen, R., Davydov, E. V., Sirota, M., & Butte, A. J. (2010). Nonsynonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One, 5*, e13574.

Chu, T. T., & Liu, Y. (2010). An integrated genomic analysis of gene-function correlation on schizophrenia susceptibility genes. *Journal of Human Genetics, 55*, 285–292.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., ... Land, S. J. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*, 80–92.

Costanzo, L. S. (2009). *Physiology*. Elsevier Health Sciences.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., ... Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry, 16*, 996–1005.

Deary, I. J., & Batty, G. D. (2007). Cognitive epidemiology. *Journal of Epidemiology and Community Health, 61*, 378–384.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*, 491–498.

Drake, J. A., Bird, C., Nemesh, J., Thomas, D. J., Newton-Cheh, C., Reymond, A., ... Hirschhorn, J. N. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics, 38*, 223–227.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics, 10*, 48.

Endo, M., Doi, R., Nishita, M., & Minami, Y. (2012). Ror family receptor tyrosine kinases regulate the maintenance of neural progenitor cells in the developing neocortex. *Journal of Cell Science, 125*(Pt 8), 2017–2029.

Gale, C. R., Deary, I. J., Boyle, S. H., Barefoot, J., Mortensen, L. H., & Batty, G. D. (2008). Cognitive ability in early adulthood and risk of 5 specific psychiatric disorders in middle age: The Vietnam experience study. *Archives of General Psychiatry, 65*, 1410–1418.

Haworth, C. M., Wright, M. J., Luciano, M., Martin, N. G., de Geus, E. J., van Beijsterveldt, C. E., ... Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry, 15*, 1112–1120.

Kerr, S. M., Campbell, A., Murphy, L., Hayward, C., Jackson, C., Wain, L. V., ... Porteous, D. J. (2013). Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Medical Genetics, 14*, 38.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., ... Christiani, D. C. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics, 91*, 224–237.

Lezak, M. (2004). *Neuropsychological testing*. Oxford, UK: Oxford University Press.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics, 25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Marth, G. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*, 2078–2079.

Liu, B., Farrington, S. M., Petersen, G. M., Hamilton, S. R., Parsons, R., Papadopoulos, N., ... Dunlop, M. G. (1995). Genetic instability occurs in the majority of young patients with colorectal cancer. *Nature Medicine, 1*, 348–352.

Luciano, M., Batty, G. D., McGilchrist, M., Linksted, P., Fitzpatrick, B., Jackson, C., ... Smith, B. H. (2010). Shared genetic aetiology between cognitive ability and cardiovascular disease risk factors: Generation Scotland's Scottish family health study. *Intelligence, 38*, 304–313.

MacLeod, A. K., Davies, G., Payton, A., Tenesa, A., Harris, S. E., Liewald, D., ... & Deary, I. J. (2012). Genetic copy number variation and general cognitive ability. *PLoS One, 7*, e37385.

Marioni, R. E., Davies, G., Hayward, C., Liewald, D., Kerr, S. M., Campbell, A., ... Deary, I. J. (2014a). Molecular genetic contributions to socioeconomic status and intelligence. *Intelligence, 44*, 26–32.

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press

Marioni, R. E., Penke, L., Davies, G., Huffman, J. E., Hayward, C., & Deary, I. J. (2014b). The total burden of rare, non-synonymous exome genetic variants is not associated with childhood or late-life cognitive ability. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140117.

McGue, M., & Christensen, K. (2013). Growing old but not growing apart: Twin similarity in the latter half of the lifespan. *Behavior Genetics, 43*, 1–12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*, 1297–1303.

McRae, A. F., Wright, M. J., Hansell, N. K., Montgomery, G. W., & Martin, N. G. (2013). No association between general cognitive ability and rare copy number variation. *Behavior Genetics, 43*, 202–207.

Meynert, A. M., Bicknell, L. S., Hurles, M. E., Jackson, A. P., & Taylor, M. S. (2013). Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics, 14*, 195.

Miller, G. F. (2000). Mental traits as fitness indicators: Expanding evolutionary psychology's adaptationism. In D. LeCroy & P. Moller (Eds.), *Evolutionary perspectives on human reproductive behavior*. Vol. *907*, (pp. 62–74). New York: New York Academy of Sciences.

Miller, G. F., & Penke, L. (2007). The evolution of human intelligence and the coefficient of additive genetic variance in human brain size. *Intelligence, 35*, 97–114.

Mouaffak, F., Kebir, O., Bellon, A., Gourevitch, R., Tordjman, S., Viala, A., ... Krebs, M. O. (2011). Association of an UCP4 (SLC25A27) haplotype with ultra-resistant schizophrenia. *Pharmacogenomics, 12*, 185–193.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., ... Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature, 461*, 272–276.

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality, 21*, 549–587.

Pisansky, M. T., Wickham, R. J., Su, J., Fretham, S., Yuan, L. L., Sun, M., ... Gewirtz, J. C. (2013). Iron deficiency with or without anemia impairs prepulse inhibition of the startle reflex. *Hippocampus, 23*, 952–962.

Posthuma, D., Luciano, M., Geus, E. J., Wright, M. J., Slagboom, P. E., Montgomery, G. W., ... Boomsma, D. I. (2005). A genomewide scan for intelligence identifies quantitative trait loci on 2q and 6p. *American Journal of Human Genetics, 77*, 318–326.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*, 559–575.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H. K. Lewis.

Smith, B. H., Campbell, H., Blackwood, D., Connell, J., Connor, M., Deary, I. J., ... Morris, A. D. (2006). Generation Scotland: The Scottish family health study; a new resource for researching genes and heritability. *BMC Medical Genetics, 7*, 74.

Smith, B. H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S. M., ... Morris, A. D. (2013). Cohort profile: Generation Scotland: Scottish family health study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology, 42*, 689–700.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*, 401–426.

Supriyanto, I., Watanabe, Y., Mouri, K., Shiroiwa, K., Ratta-Apha, W., Yoshida, M., ... Hishimoto, A. (2013). A missense mutation in the ITGA8 gene, a cell adhesion molecule gene, is associated with schizophrenia in Japanese female patients. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 40*, 347–352.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., ... Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science, 337*, 64–69.

Trzaskowski, M., Yang, J., Visscher, P. M., & Plomin, R. (2014). DNA evidence for strong genetic stability and increasing heritability of intelligence from age 7 to 12. *Molecular Psychiatry, 19*, 380–384.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 11.10.1–11.10.33.

Wechsler, D. (1997). *WAIS-III Weschler Adult Intelligence Scale*. San Antonio, TX: Psychological Coorporation.

Wechsler, D. (1998). *WMS-IIIUK administration and scoring manual*. London, UK: Psychological Corporation.

Yasuno, K., Ando, S., Misumi, S., Makino, S., Kulski, J. K., Muratake, T., ... Tamiya, G. (2007). Synergistic association of mitochondrial uncoupling protein (UCP) genes with schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 144B*, 250–253.

Yosifova, A., Mushiroda, T., Kubo, M., Takahashi, A., Kamatani, Y., Kamatani, N., ... Nakamura, Y. (2011). Genome-wide association study on bipolar disorder in the Bulgarian population. *Genes, Brain and Behavior, 10*, 789–797.

https://doi.org/10.1017/thg.2015.10 Published online by Cambridge University Press