International Journal of Obesity (2013) 37, 424–431
© 2013 Macmillan Publishers Limited All rights reserved 0307-0565/13
www.nature.com/ijo

Open

ORIGINAL ARTICLE

Determination of the obesity-associated gene variants within the entire *FTO* gene by ultra-deep targeted sequencing in obese and lean children

M Sällman Almén[1], M Rask-Andersen[1], JA Jacobsson[1], A Ameur[2], I Kalnina[3], G Moschonis[4], S Juhlin[1], N Bringeland[1], LA Hedberg[1], V Ignatovica[3], GP Chrousos[5], Y Manios[4], J Klovins[3], C Marcus[6], U Gyllensten[2], R Fredriksson[1] and HB Schiöth[1]

BACKGROUND: The *Fat mass and obesity-associated* gene (*FTO*) was the first gene reliably associated with body mass index in genome-wide association studies on a population level. At present, the genetic variations within the *FTO* gene are still the common variants that have the largest influence on body mass index.
METHODS: In the current study, we amplified the entire *FTO* gene, in total 412 Kbp, in over 200 long-range PCR fragments from each individual, from 524 severely obese and 527 lean Swedish children, and sequenced the products as two DNA pools using massive parallel sequencing (SOLiD).
RESULTS: The sequencing achieved very high coverage (median 18 000 reads) and we detected and estimated allele frequencies for 705 single nucleotide polymorphisms (SNPs) (19 novel) and 40 indels (24 novel) using a sophisticated statistical approach to remove false-positive SNPs. We identified 19 obesity-associated SNPs within intron one of the *FTO* gene, and validated our findings with genotyping. Ten of the validated obesity-associated SNPs have a stronger obesity association ($P<0.007$) than the commonly studied rs9939609 SNP ($P<0.012$).
CONCLUSIONS: This study provides a comprehensive obesity-associated variation map of *FTO*, identifies novel lead SNPs and evaluates putative causative variants. We conclude that intron one is the only region within the FTO gene associated with obesity, and finally, we establish next generation sequencing of pooled DNA as a powerful method to investigate genetic association with complex diseases and traits.

International Journal of Obesity (2013) **37,** 424–431; doi:10.1038/ijo.2012.57; published online 24 April 2012

Keywords: genetics; NGS; next generation sequencing

INTRODUCTION

Genetic variants of the *FTO* gene were the first common genetic polymorphisms to be associated with increased body weight and obesity.[1] This association was replicated in several distinct populations,[2–10] making *FTO* the single strongest contributing genetic factor to obesity, with huge implications for public health. Until now, most genetic studies have focused on variants of a 42-Kbp haplotype block, around the lead single nucleotide polymorphism (SNP) rs9939609, in the first intron. *FTO*, however, is a very large gene containing nine exons, and spanning more than 400 Kbp. Only a few studies have addressed other regions of the gene, and these have reported obesity-associated SNPs in intron two, three[11] and eight.[12] In addition, we have previously identified an SNP in intron 4, associated with insulin levels and resistance.[13]

It is unclear how *FTO* contributes to the etiology and increased susceptibility of obesity. Some reports suggest that it affects food intake, as carriers of the risk allele tend to choose high energy and palatable food.[14–16] Studies of *Fto* knockout[17] and hypomorphic mice (*Fto*^{I367A})[18] point towards a role for *Fto* in energy homeostasis, metabolism and adipogenesis, and such mice have clearly decreased body fat mass. Furthermore, transgenic mice that overexpress *Fto* have a higher body weight,[19] and in wild-type mice hypothalamic *FTO* expression is regulated in several animal-feeding models, which also suggests that *FTO* is involved in food intake and body-weight regulation.[20, 21] In agreement with the animal models, human heterozygous carriers of rs9939609 express more of the high-risk allele, suggesting that an altered *cis*-regulation site(s) within the obesity-associated haplotype of intron one causes higher *FTO* expression.[22]

Recently, new technologies have been developed that permit large-scale analysis of genomic data. These methods have lead to large biological projects, such as the ENCODE (Encyclopedia of DNA Elements) project, which aims to identify the functional elements of the human genome and provide new means for interpreting genetic data.[23] In the past few years, it has been possible to determine the genetic variability at a much higher resolution and comprehensibility than before owing to advances in massive parallel sequencing, such as the SOLiD (Sequencing by Oligonucleotide Ligation and Detection) system used in this study.[24] Nevertheless, performing individual sequencing for a

[1]Department of Neuroscience, Functional Pharmacology, Uppsala University, BMC, Uppsala, Sweden; [2]Department of Immunology, Genetics, and Pathology, SciLifeLab Uppsala, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden; [3]Latvian Biomedical Research and Study Centre, Riga, Latvia; [4]Department of Nutrition and Dietetics, Harokopio University of Athens, Athens, Greece; [5]First Department of Pediatrics, Athens University Medical School, Aghia Sophia Children's Hospital, Athens, Greece and [6]Department for Clinical Science, Intervention and Technology, Karolinska Institutet, Division of Pediatrics, National Childhood Obesity Centre, Stockholm, Sweden. Correspondence: Professor HB Schiöth, Department of Neuroscience, Functional Pharmacology, Uppsala University, Biomedical Center, Box 593, Uppsala 75124, Sweden.
E-mail: helgis@bmc.uu.se

Received 11 January 2012; revised 8 March 2012; accepted 11 March 2012; published online 24 April 2012

large number of subjects comes at a considerable cost, and processing and analyzing the large quantities of data produced from massive parallel sequencing is still at a very early stage. We have previously shown that sequencing of pooled DNA samples is a feasible and affordable way to detect novel SNPs.[25] Moreover, we have recently developed a statistical approach to model and filter away false-positive SNPs from SOLiD implemented specifically for pooled DNA samples.[26]

In the present study, we have sequenced 412 Kbp of the genome, covering the complete *FTO* gene, in 524 severely obese children and 527 lean controls with the aim to identify all common obesity-associated variations. Herein, we ascertain the first comprehensive variation map of the *FTO* gene and identify its obesity-associated variation pattern, including both SNPs and small indels.

MATERIALS AND METHODS

Collection of subjects and preparation of DNA pools

The population chosen for sequencing of the *FTO* gene comprised of a severely obese and a lean control group, both of which are sampled from the Stockholm area in Sweden. These two groups have previously been described and variations within *FTO* are known to be associated with the obese group.[27] The severely obese group consists of 524 children (270 girls and 254 boys, mean age (s.d.) 12.7 (3.2) and body mass index *z*-score (s.d.) 3.5 (0.6)) enrolled at the National Childhood Obesity Center at Karolinska University Hospital, Huddinge, and a lean group consisting of 527 healthy and normal-weight Swedish adolescents (271 girls and 256 boys, mean age (s.d.) 17.0 (0.9), mean body mass index *z*-score (s.d.) 0.06 (0.8)) recruited from 17 upper secondary schools in the Stockholm area, Sweden. The study was approved by the Regional Committee of Ethics, Stockholm, and the participants and their guardians gave their informed written consent. Genomic DNA from all subjects ($n = 1051$) was extracted from peripheral blood using QiaGen Maxiprep kit (Qiagen, Hilden, Germany). Duplicate DNA concentrations were measured with a spectrophotometer (Thermo Scientific, Waltham, MA, USA). Equal amounts of DNA from each individual were combined into pools with $170 - 190$ samples/pool, which gave three DNA pools for each group (see Figure 1 for a summary of pooling strategy and *FTO* targeting). See Supplementary Information 1 for primer design and PCR procedures.

SOLiD sequencing

Libraries were prepared from each pool and emulsion PCR was carried out according to the instructions from Applied Biosystems (Foster City, CA, USA). Sequencing was performed with Applied Biosystems' SOLiD 3 platform using a 50-bp read length on standard slides according to the protocol of the manufacturer. The reads were mapped to the *FTO* reference sequence (hg18, chr16:52 285 069-52 716 675) using the corona lite algorithm (Applied Biosystems) with default settings.

Variant calling and filtering

The methodology used for identification and filtration of false-positives SNPs, insertions and deletion (indels) are covered in greater detail in Supplementary Information 1. Briefly, SNPs are reported by the SOLiD system as valid adjacent miss-matches (that is, not sequencing errors) compared with the reference sequence. However, to limit the number of false-positive SNPs, we implemented a strategy that we previously used on data from SOLiD sequencing of pooled DNA.[26] This method focuses on filtering excessively amplified reads containing valid adjacent miss-matches, probably derived from errors in previous PCR amplification steps and not true polymorphic sites. A quality score of unique valid adjacent mismatches (UVAM) was calculated for each candidate polymorphic site and a UVAM threshold was determined by comparing the score distribution of the detected candidate sites to sites that were found in the dbSNP database and the 1000 genomes project (Figures 2a–d). Sites that had a lower quality score than the threshold was considered false positives and excluded. The filtration step was validated by genotyping 48 candidate SNPs before filtration with a 48-plex GoldenGate assay from Illumina (San Diego, CA, USA).[28]

Indels with an estimated minor allele frequency $>1\%$ were detected using two methods for gapped read mapping and included for further analysis.

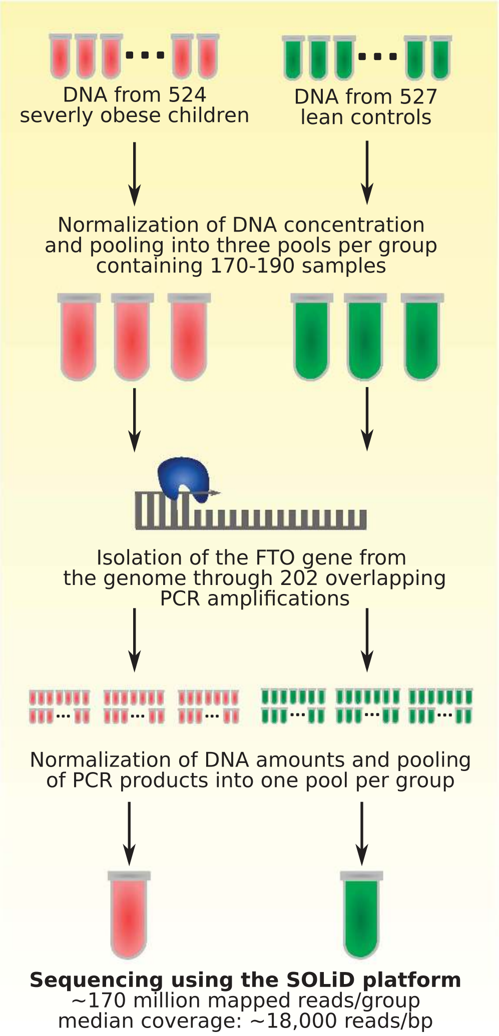
DNA from 524
severely obese children
DNA from 527
lean controls
Normalization of DNA concentration
and pooling into three pools per group
containing 170-190 samples
Isolation of the FTO gene from
the genome through 202 overlapping
PCR amplifications
Normalization of DNA amounts and pooling
of PCR products into one pool per group
Sequencing using the SOLiD platform
~170 million mapped reads/group
median coverage: ~18,000 reads/bp

Figure 1. The pooling strategy of DNA from severely obese and lean children, respectively, into two high-complex DNA pools was performed in several steps. First, each sample was normalized depending on concentration and pooled into three pools for each group. Next, specific and optimized overlapping primers for the *FTO* region were used to enrich it through 202 PCR duplicate reactions for each pool. Finally, the concentration of the PCR products were normalized and pooled into one high-complexity DNA pool for each group, which were sequenced using the SOLiD system under high coverage.

Data analysis

Association tests based on allele frequencies were performed using Fisher's exact test together with approximation of 95% confidence interval off odds ratios for each SNP and indel using the software R (http://www.r-project.org, R Foundation for statistical computing, Vienna, Austria). To adjust for multiple testing, the significance level was adjusted for the 44 haplotype blocks that we identified across the sequenced region in a previous study of a Swedish population, which contains SNPs in linkage disequilibrium.[29] Consequently, *P*-values below 1.14×10^{-3} were considered significant.

Results from the individual genotyping (GoldenGate and TaqMan assay) were analyzed with PLINK (http://pngu.mgh.harvard.edu/purcell/plink/). The validated SNPs were checked for deviation from Hardy–Weinberg equilibrium using a Pearson's χ^2-test (1 d.f) before being tested for association with obesity by using logarithmic regression to calculate 95% confidence intervals of odds ratios, controlled for sex and assuming an

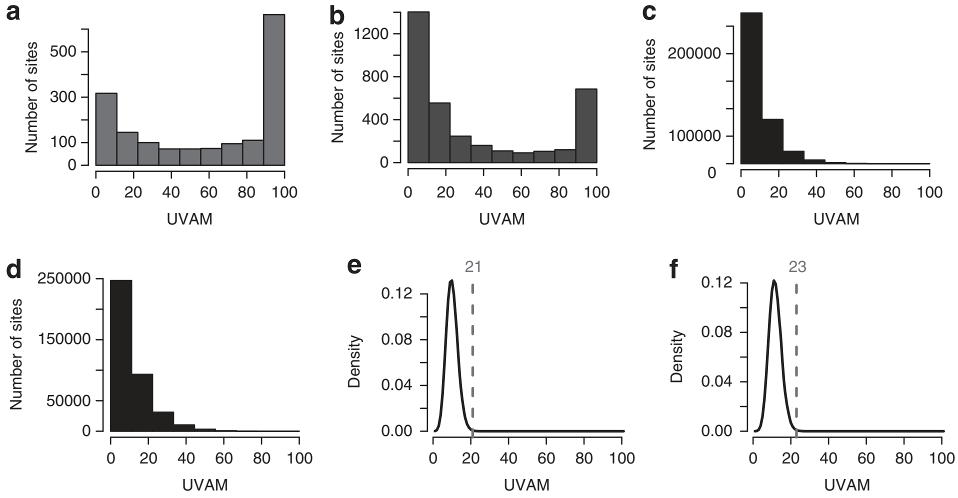

Figure 2. To remove false-positive SNPs, the distribution of unique valid adjacent miss-matches (UVAM) from the SOLiD sequencing were compared for known SNPs from the dbSNP database (**a**) and results from the 1000 genomes project (**b**) with the candidate SNPs from the sequencing of obese (**c**) and lean group (**d**). The clear difference in distribution between previously known and novel candidate SNPs indicates that most of the candidates are false positives. Based on this, a binomial hypergeometric distribution was used to model the probability of non-SNPs to have a certain UVAM score in the obese (**e**) and lean group (**f**). This allowed us to set a threshold for each group (marked in red) that corresponds to a false discovery rate of 1%. The color reproduction of this figure is available at the *International Journal of Obesity* online.

additive model. For details on the information with ENCODE data see Supplementary Information 1.

Genotyping

For details on genotyping of candidate SNPs and validation in Latvian and Greek populations see Supplementary Information 1.

RESULTS

Sequencing and mapping

In this study, we sequenced 95% of the entire human *FTO* gene and surrounding sequence, a region spanning 411 606 bases. The DNA was collected from a pool of 524 severely obese children and a pool of 527 lean controls (Figure 1). In total, we successfully sequenced 389 501 bases; the non-sequenced regions were excluded owing to difficulties in the PCR enrichment, which was primarily due to genomic repeats. After mapping, a high sequencing depth was obtained across the *FTO* region, with a median read depth of 17 385 and 18 385 in the obese and control pools, respectively. This corresponds to an ~16-fold read depth per allele in each pool for 50% of the positions (Figure 3). Furthermore, ~95% of all positions had at least 1000 in read depth in both pools, which corresponds to an average ~1-fold read depth for each FTO allele in the pools (~1000 alleles/pool) for most sites. The coverage correlates well for each position between the two pools ($r = 0.92$), and the same pattern of coverage peaks were reproduced in each independently sequenced pool (Figure 4). Characteristic peaks were found close to the primers of each amplicon, with peaks reaching a maximum of more than 345 000 fold coverage. Therefore, both primer regions, and the overlap between PCR amplicons, were excluded from further analysis. These regions add up to 40 944 bases, which leaves 348 557 bases for inclusion in the variant calling and analysis.

SNP calling and estimation of allele frequencies

Candidate SNPs were identified in reads containing valid adjacent mismatches, as defined by the SOLiD system, and allele frequencies were estimated by counting the number of reads containing the reference allele and the number of reads containing the putative SNP. To remove potential false-positive SNPs, primarily from amplification errors during the PCR enrichment, we used a statistical approach to determine a threshold for the minimum number of UVAM required for a SNP candidate to be considered reliable (Figure 2).[26] The UVAM thresholds were 21 for the obese pool and 23 for the control pool. This difference was caused by the higher coverage obtained in the control pool. In addition, SNPs were only called if they were detected in both pools, if the number of reads were distributed equally between the DNA strands (at least 23% for the strand with the lowest coverage) and if they were located outside of amplicon overlaps.

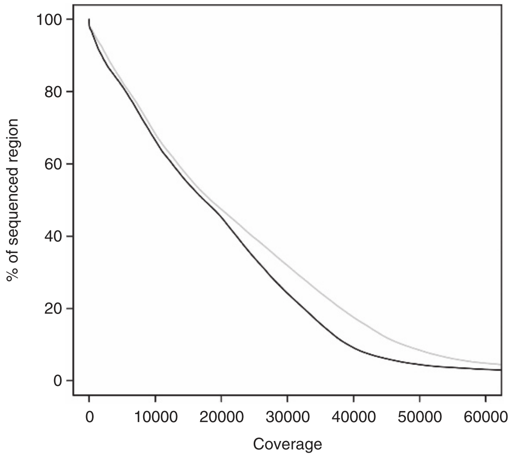

Figure 3. The graph depicts the percentage of sites that reached a certain read depth (number of reads mapped to the site). Sequencing achieved a high coverage in both groups (median ~18 000 reads). The colored lines represent the percentage of sequenced bases (*y*-axis) of the *FTO* gene in the obese (red) and lean (green) group that reach a certain coverage (*x*-axis). The color reproduction of this figure is available at the *International Journal of Obesity* online.

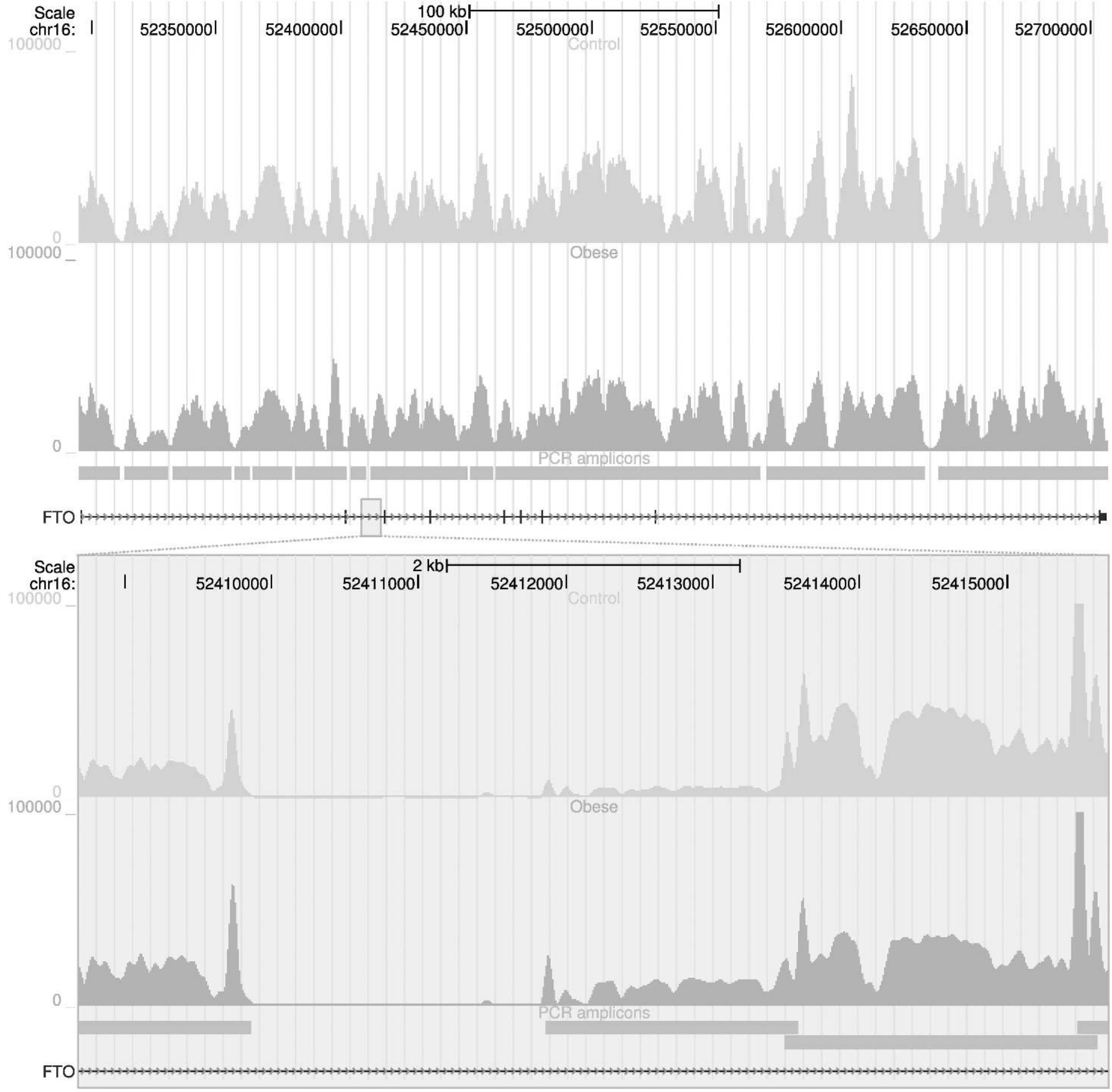

Figure 4. The *FTO* gene was sequenced at a high coverage and showed a high correlation between the lean (green) and obese (orange) pool. The *FTO* gene was targeted by overlapping specific primers to create overlapping PCR amplicons (blue). However, some regions were excluded as the PCR reaction failed, primarily due to genomic repeats. The zoom-in shows overlapping amplicons, an excluded region where the coverage drops to zero and also that the overlapping regions of the PCR products are characterized by peaks in coverage. The color reproduction of this figure is available at the *International Journal of Obesity* online.

Based on these conditions, 705 SNPs with a minor allele frequency above 1% were called, and out of these we found 19 SNPs that were not present in dbSNP(v132) and, hence, considered novel (Supplementary Information 2). In comparison, 5091 SNPs were detected when not applying the above filtration steps.

To test the validity of the UVAM filtration, we selected 48 high-frequency SNPs without first applying the UVAM threshold that had a strong obesity association for genotyping with custom-designed array. Out of the SNPs selected for genotyping, 25 were found in dbSNP, whereas the remainders were candidate novel polymorphisms. The genotyping failed for two SNPs (rs11076004 and rs3751814) and these were removed from further analysis. However, the genotyping clearly showed that all of the candidate SNPs not found in dbSNP were monoallelic, and thereby erroneously called SNPs (Supplementary Information 3). In contrast, when applying the UVAM filtration, all these false candidate SNPs were removed, which strongly validates the applied method. Among the genotyped variations found in dbSNP, four (rs11076004, rs28665159, rs34270264 and rs9928425) were found to be monoallelic. However, our filtration removed all these SNPs. Of the 17 true genotyped SNPs, four (rs1558902, rs12933996, rs9939609 and rs8063057) were removed as they were found in amplicon overlap regions or did not achieve good enough strand distribution.

The 13 genotyped true SNP correlations, between estimated allele frequencies from the SOLiD sequencing and the array had a Pearson's correlation coefficient of 0.76 ($P<2.26\times10^{-3}$) in the obese pool and 0.89 ($P<3.25\times10^{-5}$) in the control pool. Furthermore, the absolute error of the SOLiD sequencing was 5% and 8% (relative error 0.10 and 0.17) in the obese and control pool, respectively.

Indel calling and allele frequency estimation

Small insertions and deletions (indels) were called with two methods: SplitSeek, which realigns unmapped reads from corona lite; and gapped alignments using bwa and samtools. Indels were only considered valid if they were called by both approaches and frequencies, were estimated from the coverage of the alternate and reference sequence allele that were achieved by bwa. With this approach, 13 insertions and 27 deletions were called, ranging from 1 to 9 residues in length (Supplementary Information 4). Sixteen of these overlap known indels in dbSNP, whereas the remaining twenty-four are considered potential novel variants. The frequency differences between the pools were very small, with an average absolute deviation of 0.6% and a Pearson's correlation coefficient equal to 0.999.

The obesity-associated variation pattern

Fisher's exact test was performed, based on the allele frequencies for each SNP and indel with an estimated minor allele frequency above 5%, to test for association with the obese versus the lean group (Supplementary Information 2). First, 22 SNPs from the SOLiD data were considered associated with obesity ($P<1.14\times10^{-3}$) and were found in three different regions of the gene: 20 SNPs were found within the well-known obesity-associated haplotype block of intron one; 2 in intron eight and 1 in intron nine. However, after validation by genotyping with a TaqMan assay, one SNP in intron one and one in intron nine were found to be monoalleic and were excluded (Supplementary Information 3). In addition, the two SNPs in intron eight were far from association with obesity when the calculation was based on the genotyping. All of the 19 candidate obesity-associated SNPs were found within a 31 794-bp-long region (hg18 chr16:52 357 478-52 389 272) that lies within the known obesity-associated haplotype block of intron one. Ten of the 19 putative obesity-associated SNPs from the SOLiD data in intron one were genotyped with the 48-plex GoldenGate array and analyzed for association with obesity using logarithmic regression and controlling for sex (See Table 1 for a summary of the association tests for all true SNPs associated with obesity calculated on the SOLiD data). Based on the genotyping, all 10 SNPs were associated with obesity, when not adjusting for multiple testing. However, when adjusting for the 44 haplotype blocks in the *FTO* region, 3 SNPs (rs55872725, rs11642015 and rs62048402) were associated ($P<1.1\times10^{-3}$). The SNPs, rs11642015 and rs62048402, were the top-associated SNPs, when calculating association tests based on SOLiD estimates and genotyping. No indel was considered associated with obesity, even when not adjusting for multiple testing ($P>0.05$). For results from the validation in Greek and Lativan populations, see Supplementary Information 1 and 5.

By using the UCSC browser, the genetic data were integrated with ENCODE data of gene regulation and DNA sequence conservation between vertebrate species, to try to identify candidate functional variants among the obesity-associated SNPs (Figure 5 and Supplementary Information 1).

DISCUSSION

Here we provide a uniquely high resolution of the obesity-associated variation pattern of the *FTO* gene, identifying 19 obesity-associated SNPs within a haplotype block of intron one (Table 1). In total, we sequenced ~400 Kbp with very high coverage (median ~18 000 reads), and identified 705 SNPs and 40 indels across the gene (Figure 5). We conclude that it is likely that intron one is the only obesity-associated region of the *FTO* gene, as we do not detect any strongly associated variants elsewhere. None of the three SNPs, having the strongest association with obesity that we validated with genotyping, have previously been reported in any previous genetic association studies.

The SNP rs62048402 ($P<0.0007$; OR = 1.35 (confidence interval 95%: 1.13–1.60)) has a stronger association with obesity than the most studied *FTO* marker rs9939609 ($P<0.012$; OR = 1.25 (confidence interval 95%: 1.05–1.48)). Moreover, an additional nine of the SNPs that we validated have a stronger association with obesity than rs9939609 (Table 1). Intriguingly, this top obesity-associated SNP rs62048402, lies within a potential regulatory region of intron one, together with the additional three top obesity-associated SNPs (Figure 5). This region is distinguished by high DNAse sensitivity and modifications of Histone 3 (mono-methylation and acetylation), which are characteristic for accessible chromatin in enhancer sites. Further, CHiP-seq data suggests that several transcription factors bind to this region, of which the glucorticoid receptor has the strongest signal (maximum score 1000/1000, Figure 5). Thus, rs62048402, and the other SNPs, in this glucorticoid receptor-binding region could be the most likely candidate causative variants for the obesity association of the *FTO* gene, as they are the only associated SNPs in our comprehensive map, which are found within a known functional region. The potential importance of this region is emphasized by replication of the stronger association of rs62048402 and rs11642101 over rs9939609 in 2280 Greek children and 760 Latvian adults (Supplementary Information 5).

For the first time, sequencing allows us to test if small indels contribute, and could be causative, for the association of the *FTO* gene with obesity. By using two different algorithms to avoid calling false-positives, we identified 13 insertions and 27 deletions that range between 1 and 9 residues of length, of which 24 are potential novel indels. However, these indels do not seem to contribute to the obesity association of the *FTO* gene, as we did not find any association; although three insertions with a minor allele frequency below 5% reside within the obesity-associated haplotype of intron one. The average frequency for the indels is relatively low (~6% in both groups) and the allele frequency correlates almost perfectly between the pools. However, as the short read sequencing technology only allows us to identify small indels with a reasonable accuracy, it is possible that we are not detecting longer structural variants that could be associated with obesity.

The number of new SNPs that we detected can be considered to be low, as only 19 of the total number of 705 were not found in databases. This clearly suggests that recent progress in mapping of genetic variations, most notably through the 1000 genomes project, has provided a nearly complete catalog of frequent SNPs.[30] As expected, the median minor allele frequency of the novel SNPs is low (~3%). However, five of them are more common, having more than 10% in minor allele frequency. Interestingly, we only find two SNPs, over 1% minor allele frequency, in exons (both in the 5′ UTR). Hence, we conclude that mutations causing amino-acid shifts in *FTO* are very rare. This is in line with the fact that known defects in the *FTO* gene lead to gross developmental defects.[31] Moreover, our results are in good agreement with previous limited sequencing in Europeans, of *FTO* exons, that only identified rare mutations;[32] and also results in Europeans from the 1000 genomes project (http://browser.1000genomes.org/). In a perspective of the whole gene, we identified around 2 SNPs/Kbp, which can be compared with the results in Europeans of the 1000 genomes project for the whole genome with about 3 SNPs/Kbp.[30]

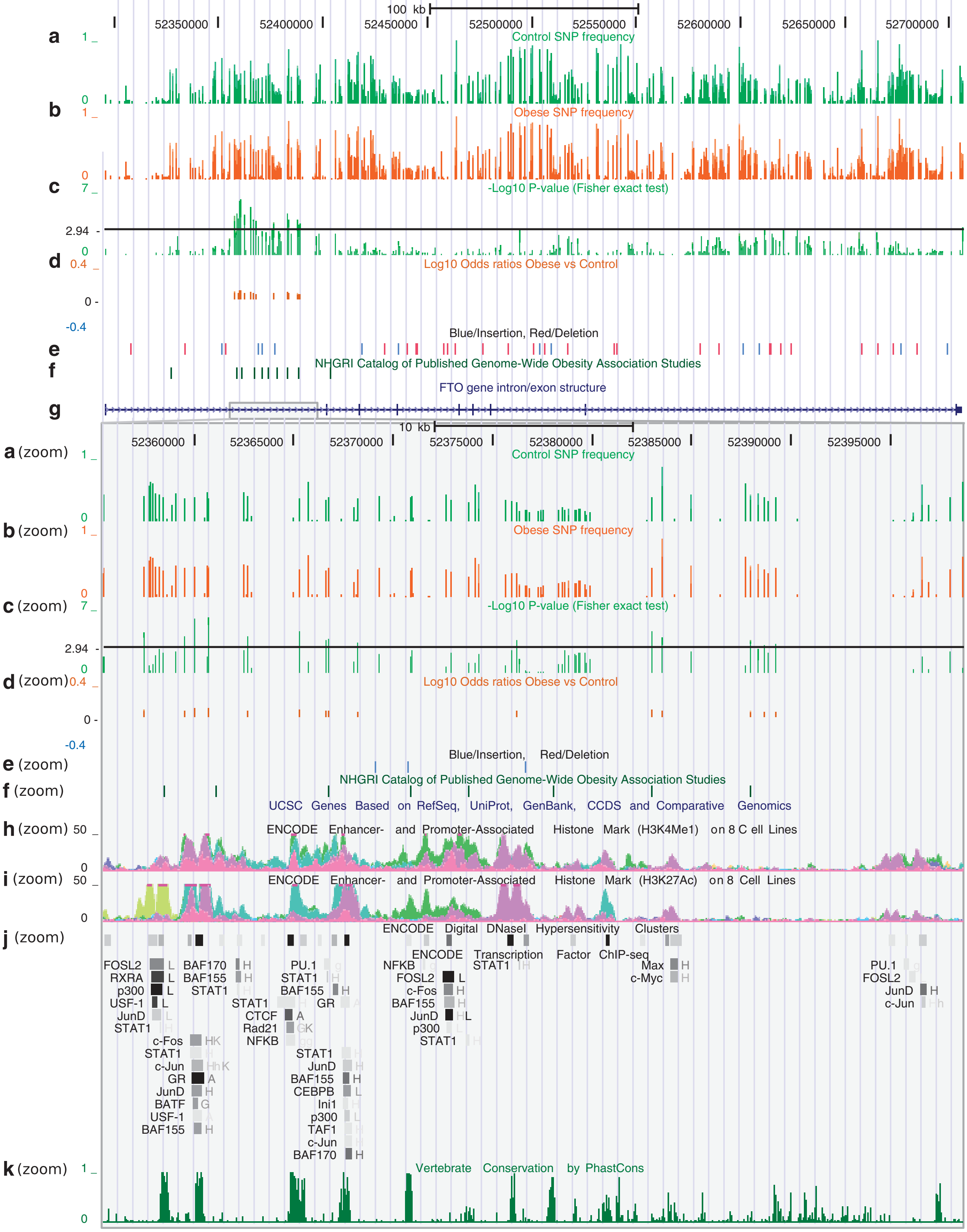

Figure 5. The genomic context of the sequenced *FTO* locus is depicted together with results from the variant calling, association tests and integration with data from the ENCODE project. The different features are represented by tracks (**a–k**) created in the UCSC browser: Track (**a**) represents the estimated alternative allele frequencies for the control group and track (**b**) for the obese group. Track (**c**) contains *P*-values from the association tests and log odds ratios are drawn in track (**d**) for those SNPs with a *P*-value$<1.14\times10^{-3}$ (black line in (**c**)). Track (**e**) represents small indels. Track (**f**) shows obesity-associated SNPs identified in previous GWASs. Track (**g**) illustrates the intron/exon structure of *FTO*. Track (**h–j**) shows data from the ENCODE project that indicates functional genomic elements using different markers (see the UCSC browser for more information). Track (**k**) shows the probability of a conserved functional region.

Table 1. Obesity-associated SNPs from the sequencing of the FTO gene

ID	*Position*	*Alternative frequency obese*		*Alternative frequency control*		P-*value*		*Odds ratio (CI: 95%)*	
		SOLiD (%)	*Genotyping (%)*	*SOLiD (%)*	*Genotyping (%)*	*SOLiD*	*Genotyping*	*SOLiD*	*Genotyping*
Intron 1									
rs9930333	52357478	48	51	39	44	5.08E-05	2.82E-03	1.43 (1.20–1.71)	1.29 (1.10–1.53)
rs9930397	52357486	46		39		3.49E-04		1.37 (1.15–1.64)	
rs9923544	52359486	44	51	36	45	9.13E-05	4.09E-03	1.42 (1.18–1.69)	1.28 (1.10–1.52)
rs11642015	52359995	48	48	37	41	1.20E-06	9.75E-04	1.54 (1.29–1.84)	1.33 (1.12–1.58)
rs1558901	52360688	43		34		2.55E-05		1.46 (1.22–1.75)	
rs62048402	52360724	46	48	35	41	1.58E-06	6.88E-04	1.54 (1.28–1.84)	1.35 (1.13–1.60)
rs11075985	52362708	48	51	40	44	1.07E-04	3.15E-03	1.41 (1.19–1.69)	1.29 (1.09–1.53)
rs17817288	52365265	44	45	35	51	2.70E-05	7.68E-03	1.46 (1.22–1.75)	0.79 (0.67–0.94)
rs55872725	52366624	41	48	33	41	3.08E-04	9.65E-04	1.39 (1.16–1.67)	1.33 (1.12–1.58)
rs1121980	52366748	46	51	37	45	1.67E-04	6.13E-03	1.40 (1.17–1.67)	1.27 (1.07–1.50)
rs7193144	52368187	43		36		5.88E-04		1.36 (1.14–1.63)	
rs9941349	52382989	46	48	37	42	3.06E-05	2.88E-03	1.45 (1.21–1.73)	1.30 (1.09–1.53)
rs28567725	52383529	42	48	35	42	2.36E-04	3.25E-03	1.39 (1.16–1.67)	1.29 (1.09–1.53)
rs9930501	52387953	36		29		1.12E-03		1.36 (1.13–1.64)	
rs9930506	52387966	39		31		2.95E-04		1.39 (1.16–1.67)	
rs9932754	52387992	48		41		8.36E-04		1.35 (1.13–1.61)	
rs9922708	52388647	44		36		7.21E-04		1.35 (1.13–1.61)	
rs9922619	52389272	44		37		6.21E-04		1.36 (1.14–1.62)	
Intron 8									
rs11860076	52515976	53	44	44	48	4.06E-05	6.28E-02	1.44 (1.21–1.71)	0.85 (0.71–1.01)
rs11864972	52520564	21	35	29	36	4.59E-05	0.6214	0.66 (0.54–0.81)	0.96 (0.80–1.14)

Abbreviation: CI, confidence interval.

Targeted large-scale sequencing of many individuals simultaneously, in high-complexity DNA pools, is a new cost-effective approach to provide a high-resolution variation map. We have clearly demonstrated that sequencing of pooled DNA from different groups gives high enough accuracy to detect associated common variants in complex traits, providing appropriate data management. We show that our filtration approach, which relies on UVAM scores and additional strict filters, while retaining all true SNPs, was able to appropriately exclude all 20 false positives that were erroneously called using a tolerant approach, relying mainly on the SOLiD systems definition of valid SNPs. This clearly shows that data from massive parallel sequencing technologies demands careful and specific implementation of data processing and analysis to give reliable results. Interestingly, we identified two SNPs (rs2388452 and rs1421086), found in dbSNP and the 1000 genomes project, which were not excluded by the UVAM filtration, but are monoallelic sites according to the TaqMan assay. Assuming that the traditional genotyping is reliable, this could indicate that the novel sequencing technologies used in this study and the 1000 genome project may cause reproducible artifacts that erroneously are called as SNPs. Although powerful, this suggests that massive parallel sequencing technologies needs to be more extensively compared with traditional genotyping methods to prove their definite accuracy.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The study was supported by the Swedish Research Council, the Brain Research Foundation, Novo Nordisk Foundation and the Åhlens Foundation. We thank Dr Michael J Williams for valuable comments on the article. Genotyping was performed by the SNP&SEQ Technology Platform in Uppsala, which is supported by Uppsala University, Uppsala University Hospital, Science for Life Laboratory - Uppsala and the Swedish Research Council (Contracts 80576801 and 70374401). The SOLiD sequencing was performed by the Uppsala Genome Center that is part of the Science for Life Laboratory Uppsala Genomics Platform. We thank the Healthy Growth Study Group for their contribution in this study. The Healthy Growth Study Group consists of (1) Harokopio University Research Team/Department of Nutrition and Dietetics: Yannis Manios (Coordinator), George Moschonis (Project manager), Katerina P Skenderi, Evangelia Grammatikaki, Odysseas Androutsos, Sofia Tanagra, Alexandra Koumpitski, Paraskevi-Eirini Siatitsa, Anastasia Vandorou, Aikaterini-Efstathia Kyriakou, Vasiliki Dede, Maria Kantilafti, Aliki-Eleni Farmaki, Aikaterini Siopi, Sofia Micheli, Louiza Damianidi, Panagiota Margiola, Despoina Gakni, Vasiliki Iatridi, Christina Mavrogianni, Kelaidi Michailidou, Aggeliki Giannopoulou, Efsthoula Argyri, Konstantina Maragkopoulou, Maria Spyridonos, Eirini Tsikalaki, Panagiotis Kliasios, Anthi Naoumi, Konstantinos Koutsikas, Katerina Kondaki, Epistimi Aggelou, Zoi Krommyda, Charitini Aga, Manolis Birbilis, Ioanna Kosteria, Amalia Zlatintsi, Elpida Voutsadaki, Eleni-Zouboulia Papadopoulou, Zoi Papazi, Maria Papadogiorgakaki, Fanouria Chlouveraki, Maria Lyberi, Nora Karatsikaki-Vlami, Eva Dionysopoulou and Efstratia Daskalou; (2) Aristotle University of Thessaloniki/School of Physical Education and Sports Sciences: Vassilis Mougios, Anatoli Petridou, Konstantinos Papaioannou, Georgios Tsalis, Ananis Karagkiozidis, Konstantinos Bougioukas, Afroditi Sakellaropoulou and Georgia Skouli; (3) University of Athens/ Medical School: George P Chrousos, Maria Drakopoulou and Evangelia Charmandari.

REFERENCES

1 Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**: 889–894.

2 Bollepalli S, Dolan LM, Deka R, Martin LJ. Association of FTO gene variants with adiposity in African-American adolescents. *Obesity* 2010; **18**: 1959–1963.

3 Dina C, Meyre D, Gallina S, Durand E, Korner A, Jacobson P *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* 2007; **39**: 724–726.

4 Grant SF, Li M, Bradfield JP, Kim CE, Annaiah K, Santa E *et al.* Association analysis of the FTO gene with obesity in children of Caucasian and African ancestry reveals a common tagging SNP. *PLoS One* 2008; **3**: e1746.

5 Hennig BJ, Fulford AJ, Sirugo G, Rayco-Solon P, Hattersley AT, Frayling TM *et al.* FTO gene variation and measures of body mass in an African population. *BMC Med Genet* 2009; **10**: 21.

6 Hotta K, Nakata Y, Matsuo T, Kamohara S, Kotani K, Komatsu R *et al.* Variations in the FTO gene are associated with severe obesity in the Japanese. *J Hum Genet* 2008; **53**: 546–553.

7 Peeters A, Beckers S, Verrijken A, Roevens P, Peeters P, Van Gaal L *et al*. Variants in the FTO gene are associated with common obesity in the Belgian population. *Mol Genet Metab* 2008; **93**(4): 481–484.

8 Renstrom F, Payne F, Nordstrom A, Brito EC, Rolandsson O, Hallmans G *et al*. Replication and extension of genome-wide association study results for obesity in 4923 adults from northern Sweden. *Hum Mol Genet* 2009; **18**(8): 1489–1496.

9 Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J *et al*. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 2007; **3**: e115.

10 Xi B, Wang C, Wang R, Huang Y. FTO gene polymorphisms are associated with obesity and type 2 diabetes in East Asian populations: an update. *Obesity* 2011; **19**: 236–237; author reply 238.

11 Tonjes A, Zeggini E, Kovacs P, Bottcher Y, Schleinitz D, Dietrich K *et al*. Association of FTO variants with BMI and fat mass in the self-contained population of Sorbs in Germany. *Eur J Hum Genet* 2010; **18**: 104–110.

12 Adeyemo A, Chen G, Zhou J, Shriner D, Doumatey A, Huang H *et al*. FTO genetic variation and association with obesity in West Africans and African Americans. *Diabetes* 2010; **59**: 1549–1554.

13 Jacobsson JA, Klovins J, Kapa I, Danielsson P, Svensson V, Ridderstrale M *et al*. Novel genetic variant in FTO influences insulin levels and insulin resistance in severely obese children and adolescents. *Int J Obes* 2008; **32**: 1730–1735.

14 Speakman JR, Rance KA, Johnstone AM. Polymorphisms of the FTO gene are associated with variation in energy intake, but not energy expenditure. *Obesity* 2008; **16**: 1961–1965.

15 Tanofsky-Kraff M, Han JC, Anandalingam K, Shomaker LB, Columbo KM, Wolkoff LE *et al*. The FTO gene rs9939609 obesity-risk allele and loss of control over eating. *Am J Clin Nutr* 2009; **90**: 1483–1488.

16 Wardle J, Carnell S, Haworth CM, Farooqi IS, O'Rahilly S, Plomin R. Obesity associated genetic variation in FTO is associated with diminished satiety. *J Clin Endocrinol Metab* 2008; **93**: 3640–3643.

17 Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Bruning JC *et al*. Inactivation of the Fto gene protects from obesity. *Nature* 2009; **458**: 894–898.

18 Church C, Lee S, Bagg EA, McTaggart JS, Deacon R, Gerken T *et al*. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS Genet* 2009; **5**: e1000599.

19 Church C, Moir L, McMurray F, Girard C, Banks GT, Teboul L *et al*. Overexpression of Fto leads to increased food intake and results in obesity. *Nat Genet* 2010; **42**: 1086–1092.

20 Fredriksson R, Hagglund M, Olszewski PK, Stephansson O, Jacobsson JA, Olszewska AM *et al*. The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain. *Endocrinology* 2008; **149**: 2062–2071.

21 Olszewski PK, Fredriksson R, Olszewska AM, Stephansson O, Alsio J, Radomska KJ *et al*. Hypothalamic FTO is associated with the regulation of energy intake not feeding reward. *BMC Neurosci* 2009; **10**: 129.

22 Berulava T, Horsthemke B.. The obesity-associated SNPs in intron 1 of the FTO gene affect primary transcript levels. *Eur J Hum Genet* 2010; **18**: 1054–1056.

23 Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**: 799–816.

24 Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009; **55**: 641–658.

25 Ingman M, Gyllensten U. SNP. Frequency estimation using massively parallel sequencing of pooled DNA. *Eur J Hum Genet* 2009; **17**: 383–386.

26 Zaboli G, Ameur A, Igl W, Johansson A, Hayward C, Vitart V *et al*. Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits. *Eur J Hum Genet* 2012; **20**: 77–83.

27 Jacobsson JA, Danielsson P, Svensson V, Klovins J, Gyllensten U, Marcus C *et al*. Major gender difference in association of FTO gene variant among severely obese children with obesity and obesity related phenotypes. *Biochem Biophys Res Commun* 2008; **368**: 476–482.

28 Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL *et al*. Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* 2003; **68**: 69–78.

29 Jacobsson JA, Almen MS, Benedict C, Hedberg LA, Michaelsson K, Brooks S *et al*. Detailed analysis of variants in FTO in association with body composition in a cohort of 70-year-olds suggests a weakened effect among elderly. *PLoS One* 2011; **6**: e20158.

30 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.

31 Boissel S, Reish O, Proulx K, Kawagoe-Takaki H, Sedgwick B, Yeo GS *et al*. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. *Am J Hum Genet* 2009; **85**: 106–111.

32 Meyre D, Proulx K, Kawagoe-Takaki H, Vatin V, Gutierrez-Aguilar R, Lyon D *et al*. Prevalence of loss-of-function FTO mutations in lean and obese individuals. *Diabetes* 2010; **59**: 311–318.

Supplementary Information accompanies the paper on International Journal of Obesity website (http://www.nature.com/ijo)