ORIGINAL ARTICLE

Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia

Marta Futema,[1] Vincent Plagnol,[2] Ros A Whittall,[1] H Andrew W Neil,[3] on behalf of the Simon Broome Register Group, Steve Eric Humphries,[1] UK10K[4]

▶ Additional data are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/jmedgenet -2012-101189)

[1]Centre for Cardiovascular Genetics, British Heart Foundation Laboratories, Institute of Cardiovascular Science, The Rayne Buliding University College London, London, UK
[2]Department of Genetics, Environment and Evolution, UCL Genetics Institute, University College London, London, UK
[3]Department of Primary Care Health Sciences, NIHR School of Primary Care Research, University of Oxford, Oxford, UK
[4]http://www.uk10k.org

Correspondence to
Professor Steve Eric Humphries, Centre for Cardiovascular Genetics, British Heart Foundation Laboratories, Institute Cardiovascular Science, University College London Medicine School, The Rayne Building, 5 University Street, London WC1E 6JF, UK; rmhaseh@ucl.ac.uk

Received 25 July 2012
Revised 2 September 2012
Accepted 4 September 2012

ABSTRACT

Background Familial Hypercholesterolaemia (FH) is an autosomal dominant disease, caused by mutations in *LDLR*, *APOB* or *PCSK9*, which results in high levels of LDL-cholesterol (LDL-C) leading to early coronary heart disease. An autosomal recessive form of FH is also known, due to homozygous mutations in *LDLRAP1*. This study assessed the utility of an exome capture method and deep sequencing in FH diagnosis.

Methods Exomes of 48 definite FH patients, with no mutation detected by current methods, were captured by Agilent Human All Exon 50Mb assay and sequenced on the Illumina HiSeq 2000 platform. Variants were called by GATK and SAMtools.

Results The mean coverage of FH genes varied considerably (*PCSK9*=23x, *LDLRAP1*=36x, *LDLR*=56x and *APOB*=93x). Exome sequencing detected 17 *LDLR* mutations, including three copy number variants, two *APOB* mutations, missed by the standard techniques, two *LDLR* novel variants likely to be FH-causing, and five *APOB* variants of uncertain effect. Two variants called in *PCSK9* were not confirmed by Sanger sequencing. One heterozygous mutation was found in *LDLRAP1*.

Conclusions High-throughput DNA sequencing demonstrated its efficiency in well-covered DNA regions, in particular *LDLR*. This highly automated technology is proving to be effective for heterogeneous diseases and may soon replace laborious conventional methods. However, the poor coverage of gene promoters and repetitive, or GC-rich sequences, remains problematic, and validation of all identified variants is currently required.

INTRODUCTION

Familial Hypercholesterolaemia (FH) is a common autosomal dominant genetic disease caused by mutations affecting the plasma clearance of LDL-cholesterol (LDL-C).[1] FH patients have elevated levels of total and low-density lipoprotein (LDL) from birth, and if untreated, develop coronary heart disease (CHD) by the age of 55 in 50% of men and 30% of women.[2] In addition to the increased LDL-C, a proportion of FH patients is characterised by the occurrence of tendon xanthomas (TX), and the UK Simon Broome criteria[3] classifies TX-positive patients as Definite FH (DFH), and TX-negative patients as Possible FH (PFH), with the DFH group having a three times higher risk of developing CHD when compared with the PFH subjects.[4 5] Statin therapy has been proven highly effective in the treatment of FH patients, and the importance of an early identification of FH patients for the statin treatment has been demonstrated[6] and recently highlighted.[7]

The clinical phenotype of FH is known to be due to mutations in three genes encoding proteins involved in the clearance of LDL-C from the plasma, *LDLR*, *APOB* and *PCSK9*. There are over 1200 different *LDLR* mutations,[8] but only one common *APOB* (c.10580G>A, p.R3527Q) and one *PCSK9* (c.1120G>T, p.D374Y) mutation, reported in the UK population.[9] The majority of pathogenic *LDLR* variants are single nucleotide changes leading to significant alterations in the amino acid sequence of the mature protein, or creation of a truncated peptide. FH is also caused by variants that affect correct splicing, and by changes in the transcription-factor-binding elements located in the promoter.[10]

While *LDLR/APOB/PCSK9* mutations cause a dominant pattern of inheritance, an autosomal recessive hypercholesterolaemia (ARH) has also been observed. The locus for ARH was mapped to a chromosome 1 gene, the *LDLRAP1*, in which both homozygous and compound heterozygous mutations can be found.[11] Most of the ARH-causing mutations are due to premature stop codons.

In 2008, the UK National Institute for Health and Clinical Excellence (NICE) guidelines were published, and included a recommendation that all FH patients be offered a DNA test to confirm their diagnosis and, so that mutation confirmation could be used, to cascade-test their first-degree relatives. Newly identified patients can then be offered statin treatment.[12] In most laboratories, FH mutation screening includes use of commercially available kits designed to test for the most common mutations, such as Elucigene FH20 (Gen-Probe Life Sciences, UK) and LIPOchip (Progenica Biopharma, Spain), and for large gene rearrangements (deletions or duplications), which account for 4%–5% of all FH mutations.[13] However, in the UK, due to the highly heterogeneous nature of the population this approach is not fully effective, and many patient samples require screening the *LDLR* promoter and coding regions, splice sites and splice branch points for causative mutations, and in the diagnostic laboratory this is currently performed using Sanger sequencing. Because of the time and labour of these methods, there has been interest in next-generation sequencing (NGS) technology for the diagnosis of genetic disorders. However, whether NGS is ready for clinical use has been questioned.[14] Main limitations of the technology include the requirement for complex data analysis,

significant computing infrastructure with respect to data analysis and storage, and legal and ethical issues associated with incidental findings from acquiring whole exome data. In the research laboratory a four-phased approach is used to screen FH patients to identify the causative mutation, using the commercially available Amplification Refractory Mutation System kit, which tests for 20 of the most common UK mutations, followed by High-Resolution Melting (HRM) to detect changes within the coding region and splice sites of FH genes, followed by Multiplex Ligation-dependent Probe Amplification (MLPA), for the detection of large *LDLR* gene rearrangements, and finally Sanger sequencing.[15–17] Using standard molecular diagnostic techniques, an FH-causing mutation can be detected in 20%–30% of PFH patients, and in 60%–80% of DFH patients.[18]

The UK10K is a large-scale deep-sequencing project, based on collaboration between multiple investigators at the Wellcome Trust Sanger Institute, and clinical experts in different genetic diseases. A total of 125 FH samples with no *LDLR/APOB/PCSK9* mutation are currently in the exome sequencing pipeline, as a part of the Rare Diseases group of the UK10K. The aim of the project is to provide collaborators with high-quality exome data, which will be used for the discovery of novel disease genes. This paper reports the sequencing results of the first 48 FH exomes, and discusses sensitivity problems of the current FH mutation-screening methods, as well as demonstrating advantages and limitations of the whole exome sequencing approach.

MATERIALS AND METHODS

Patients' selection

Forty-eight unrelated FH patients were selected from the Simon Broome FH register.[19] All individuals were Caucasian and attended a lipid clinic in London, Oxford or Manchester. Patients were diagnosed using the UK Simon Broome criteria as DFH on the basis of the presence or history of TX. The entire promoter and coding regions, including splice sites, of the *LDLR* gene were screened by the HRM method, as previously described,[16] on the Rotor-Gene (6000) real-time rotary analyser. Patients were screened for presence of the *APOB* mutation, p.(R3527Q), using a restriction enzyme digest,[20] and the entire coding region of the *PCSK9* was examined by HRM.[17] Fragments with a heterozygous melting curve were analysed further by direct sequencing. Screening for large rearrangements within the *LDLR* gene was done using the MLPA[21] SALSA P062 *LDLR* kit from MRC-Holland (Amsterdam). One hundred and ninety five non-FH Caucasian samples, sequenced in parallel with the FH cohort, as a part of the UK10K rare disease arm project (http://www.uk10k.org/studies/rarediseases.html), were used as controls. None of these subjects had disorders known to affect plasma lipid levels.

Whole exome sequencing

Genomic DNA (1–3 μg), extracted from blood,[22] was sheared to 100–400 bp using a Covaris E210 or LE220 (Covaris, Woburn, Massachusetts, USA). Sheared DNA was subjected to Illumina paired-end DNA library preparation and enriched for target sequences (Agilent Technologies, Santa Clara, CA, USA; Human All Exon 50 Mb - ELID S02972011) according to the manufacturer's recommendations (Agilent Technologies, Santa Clara, CA, USA; SureSelectXT Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing). Enriched libraries were sequenced (eight samples over two lines) using the HiSeq 2000 platform (Illumina) as paired-end 75 base reads according to the manufacturer's protocol.

Bioinformatic analysis

To improve raw alignment BAMs for single nucleotide polymorphism (SNP) calling, we realigned around known (1000 Genomes pilot) indels, and recalibrated base quality scores using GATK. BAQ tags were added using samtools calmd. BAMs were merged to sample level and duplicates marked using Picard. Variants (SNPs and indels) were called on each sample individually with both samtools mpileup (0.1.17) and GATK UnifedGenotyper (1.3–21), restricted to exon bait regions plus or minus a 100 bp window. Various quality filters were applied to each of the callsets separately. Calls were then merged, giving preference to GATK information when possible. Calls were annotated with 1000 Genomes allele frequencies, dbSNP132 rsIDs and earliest appearance in dbSNP. Functional annotation was added using Ensembl Variant Effect Predictor v2.2 against Ensembl 64, and included coding consequence predictions, Sorting Tolerant From Intolerant (SIFT), PolyPhen and Condel annotations, and Genomic Evolutionary Rate Profiling (GERP) and Grantham Matrix scores. Variants previously reported by the 1000 Genomes project with minor allele frequency higher than 0.01 were filtered out. Variants that passed the initial filtering were compared against 195 non-FH control whole exomes, processed using the same pipeline, and only FH unique changes were further assessed. Pathogenicity of any private (ie, specific to the FH cohort) variant was examined using previous knowledge and by bioinformatic mutation-prediction tools, which included: PolyPhen2, SIFT, Condel, Mutation Taster, PhyloP and Gratham Score algorithm. Sanger sequencing was used to confirm presence of any identified predicted pathogenic variant.

Copy number variants analysis

The copy number variants (CNV) analysis uses a read depth strategy designed to overcome biases associated with sequence capture and high-throughput sequencing. This set of tools is implemented in the package ExomeDepth (freely available at the Comprehensive R Archive Network.[23]

RESULTS

Overall, the mean read depth for the whole exome sequence was 72x, with 78.9% of the exome covered at least 20x; and 55.8% of the targeted sequence was covered 50x or more.

LDLR analysis

The average read depth of *LDLR* exons varied from 136x for exon 12, to 4x for exon 18 (figure 1). Using a 16x coverage threshold, which would give a 99% probability of observing a rare allele at least three times, all except exons 1 and 18 showed adequate coverage. Exons 3 and 4 contain the largest number of reported FH-causing mutations,[24] and both these exons were well covered (mean depth 92 and 57, respectively).

As shown in table 1, in 14 out of the 48 samples, a FH-unique variant in the *LDLR* was called, in 11 of these the variant has been previously reported to be FH-causing.[25]

All these variants were confirmed to be correct by Sanger sequencing of duplicate DNA samples (not shown). The variants included five different missense mutations and five nonsense mutations. The two novel *LDLR* variants included: c.695-6_698del and c.1776_1778del (p.(G592del)). The c.695-6_698del is predicted to cause a frameshift and a premature stop codon by altering *LDLR* splicing (see online supplementary figure S1). The deletion of Glycine at residue 592 is predicted to disrupt packaging of the LDL-R propeller blades in the epidermal growth factor (EGF) domain, which could affect displacement of the

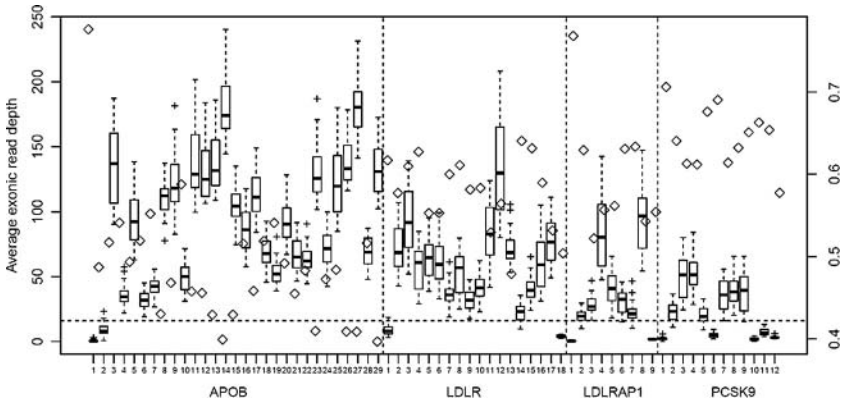


Figure 1 The exonic coverage of FH genes: *APOB*, *LDLR*, *LDLRAP1* and *PCSK9* in a standard box plot (the minimum, lower quartile, median, upper quartile and maximum for each gene exon). The horizontal dashed line indicates the 16x coverage, when the probability of observing a rare allele at least 3 times is 99%. The additional Y-axis describes GC content (0.4=40%, 0.5=50%, etc) for a given exon, shown as rhombuses.

ligand from the ligand-binding region (see online supplementary figure S2). Of these 14 mutations, all should have been detected by our standard screening protocol described in the Materials and methods section, except for c.695-6_698del, where the change was located in the primer sequence used for PCR.

CNV calling identified one deletion of exons 11 and 12 (c.1587−?_1845+?del), and two duplications of exons 3–8 (191−?_1186+?dup), and exons 13–15 (c.1846−?_2311+?dup), as shown in figure 2. All CNVs were confirmed by MLPA (see online supplementary figure S3).

APOB analysis

The mean read depth of *APOB* exons was 93. Exons 26 and 29 were covered on average 135x, whereas exon 1 was covered only once (figure 1). Two individuals were found to carry the FH-causing *APOB* mutation in exon 26, the p.(R3527Q). There were no nonsense or frameshift mutations observed in the gene sequence. One novel non-synonymous variant was found in exon 26 of *APOB*, the p.(A3426V), which was unique to the FH cohort. The variant was predicted as 'Tolerated', 'Benign' and 'Polymorphism' by SIFT, PolyPhen and Mutation Taster, respectively. Four other FH-unique non-synonymous variants were observed outside of the ligand-binding domain (see online supplementary table S1). The functional impact of these variants, as predicted by PolyPhen/SIFT/Mutation Taster, was not consistent, and whether or not these are FH-causing is unclear. There were no CNVs found within the *APOB* gene.

PCSK9 analysis

The mean read depth of the *PCSK9* exons was 23. Of these, only 58% of the gene coding sequence had the mean coverage higher than 16, whereas exons 1, 6, 10, 11 and 12 were covered 4x on average (figure 1). Exon 7, where the common UK FH-causing mutation (c.1120G>T, p.(D374Y)) occurs, was covered 36x. Two novel non-synonymous variants were called by the exome sequencing, c.1027G>C and c.1028A>C, both present in the same sample. However, despite the high read depth (51x and 50x), and the high number of read count for the novel alleles (19 and 26) (see online supplementary figure S4) the Sanger sequencing did not confirm the variants. There were no CNVs observed in *PCSK9*.

LDLRAP1 analysis

The average read depth of *LDLRAP1* was 36 with all, except exons 1 and 9 covered above the 16x threshold (figure 1). The *LDLRAP1* variant analysis was performed using a homozygosity-based strategy, and the presence of compound heterozygote variants was also assessed. There were no homozygous or compound heterozygous functional changes within the gene in any of the individuals. One patient was found to be heterozygous for a known Sicilian/Sardinian ARH mutation, the c.432_433insA, p.(A145KfsX26),

Table 1 Summary of pathogenic single nucleotide changes and small deletions/insertions in the FH genes

Gene	Samples (n)	Nucleotide change	Functional effect	Depth	Quality	Comments
LDLR	1	c.326G>A	p.(C109Y)	43	506	known FH mutation
	1	c.1690A>C	p.(N564H)	36	343	known FH mutation
	1	c.1823C>T	p.(P608L)	82	1214	known FH mutation
	1	c.2054C>T	p.(P685L)	20	135	known FH mutation
	1	c.2479G>A	p.(V827I)	65	749	known FH mutation
	2	c.682G>T	p.(E228X)	13	155	known FH mutation
	1	c.1048C>T	p.(R350X)	60	816	known FH mutation
	1	c.1150C>T	p.(Q384X)	20	275	known FH mutation
	1	c.1685G>A	p.(W562X)	41	701	known FH mutation
	1	c.2140+1G>A	Splicing	22	258	known FH mutation
	1	c.695-6_698del	Splicing	36	1543	novel
	2	c.1776_1778del	p.(G592del)	148	2634	novel
APOB	1	c.10277G>A	p.(A3426V)	192	2785	novel
	2	c.10580C>T	p.(R3527Q)	161	2144	known FH mutation
PCSK9	1	c.1027G>C	p.(D343H)	51	44	false positive
	1	c.1028A>C	p.(D343A)	50	201	false positive
LDLRAP1	1	c.432_433insA	p.(A145KfsX26)	90	2186	heterozygous

'Depth' refers to the coverage depth; 'Quality' values are Phred-like quality scores generated by SAMtools.

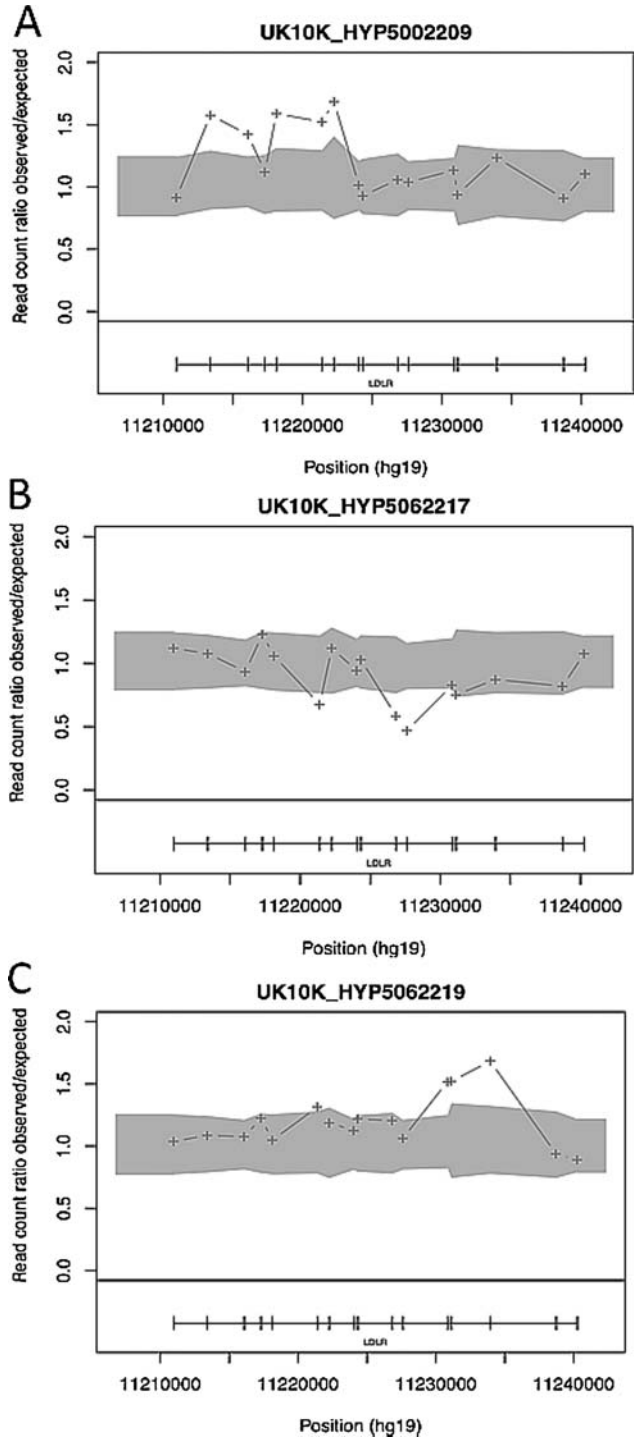


Figure 2 Copy Number Variants (CNVs) in *LDLR* gene. A: heterozygous duplication of exons 3–8. B: heterozygous deletion of exons 11 and 12. C: heterozygous duplication of exons 13–15. All identified by ExomeDepth in the exome sequencing data. The crosses show the ratio of observed/expected number of reads for the test sample. The grey shaded region shows the estimated 99% CI for this observed ratio in the absence of CNV call. The presence of contiguous exons with read count ratio located outside of the CI is indicative of a heterozygous deletion or duplication in a sample. Exons 1 and 18 were excluded from the analysis (not shown on the graph) as they did not reach the threshold of 100 for the total number of reads. All CNVs were confirmed by MLPA experiment (see online supplementary figure S3).

which is a frameshift mutation resulting in a truncated peptide formation.[11 26] Further analysis of this sample showed no other pathogenic variants in known FH genes, which could contribute to the phenotype. CNV calling did not detect large rearrangements in the *LDLRAP1*.

DISCUSSION

Current FH-screening methods

The exome sequencing results exposed sensitivity problems with the current FH mutation-screening methods used in our research laboratory. Overall, the standard variant-detection process already in place (HRM, MLPA and Sanger sequencing) did not detect 17 *LDLR* mutations (including 3 CNVs) and 2 *APOB* mutations. Although the HRM has proved to be efficient at detecting FH variants,[16] its sensitivity decreases in some gene regions, depending on the nucleotide composition of the fragment. Re-examining previous results for the samples with a *LDLR* or *APOB* variant called by the NGS, we observed that most of the variants showed a melting curve shift during the HRM assay, but Sanger sequencing of the identified gene region did not detect any heterozygous changes in the sequence despite being repeated several times (i.e. only the predicted wild-type sequence was obtained). After the exome sequencing, the Sanger sequencing was repeated on a duplicate DNA sample, and the predicted mutations were confirmed to be present, validating the exome sequencing and variant calling. Although Sanger sequencing is considered to be the gold standard mutation-detection method, a combination of PCR artefacts and the human error aspect in the protocol appears to be the main reason for the false negative calling in the original screening.

Novel *LDLR* variants

Two novel variants in *LDLR* were identified, a deletion of 10 bp on boundary of intron 4 and exon 5, which is predicted to cause a frameshift resulting in a premature stop codon by altering *LDLR* splicing, and a three bp deletion which deletes Glycine at residue 592, which is predicted to disrupt packaging of the LDL-R propeller blades in the EGF domain. Neither of these variants are found in dbSNP, the 1000 Genomes or the NHLBI Exome Sequencing Project, and are highly likely to be FH-causing, although further work is required to confirm this.

Novel *APOB* variants

APOB codes for one of the largest human proteins, which is the major component of the LDL-C responsible for binding to the LDL receptor.[27] The actual binding site for the receptor, the B-site (residues 3386–3396), has been mapped to a region encoded by exon 26 of the *APOB*, which is the longest coding exon known (7572 bps).[28] In addition, the C-terminus encoded by exon 29 of the gene was proposed to function as a modulator of the receptor binding.[28] Therefore, our variant analysis strategy prioritised novel variants located in exons 26 and 29 of the gene, as these are more likely to cause the FH phenotype. In this study, there was only one novel variant identified in the exon 26 of *APOB*, the c.10277G>A (p.(A3426V)), which was not observed previously by the dbSNP, the 1000 Genomes or the NHLBI Exome Sequencing Project. The variant was not present in the 195 non-FH exomes from the UK10K project, which were processed using the same pipeline, increasing the likelihood that it is in fact disease-causing. The novel p.

(A3426V) variant is located near to the LDL receptor-binding site (B-site), and close to the known FH mutation p.(R3527Q), and although it does not alter the charge at the site, it may produce a conformational change affecting the LDL-R/ApoB interaction. This requires further experiments since the current in silico prediction tools are not able to assess protein–protein interactions. We will also examine whether or not the variant cosegregates with the disease. Four other novel *APOB* variants were identified in this group of patients in the N-terminal part of the protein. Although these variants are less likely to influence LDL clearance from the blood, since the N-terminal region of the protein is not involved in interacting with the LDL-R, some of the variants are predicted as damaging by Polyphen or SIFT. The aim of this study was to assess the clinical utility of exome sequencing as a sensitive mutation detection tool, rather than finding novel FH mutations. Future work includes the assessment of novel identified variants, which will involve family cosegregation and functional assays.

Promoter region analyses

Most of the sequencing data generated for Mendelian disorders are focused on the exome, which constitutes around 1% of the whole human genome. Prediction tools for the analysis of non-synonymous changes are well established and widely used to estimate the deleteriousness of amino acid changes. However, since the majority of human variations are located in the non-coding regions,[29] concern about the bias towards variants in the protein-coding sequence was highlighted.[30] Proving the pathogenic effect of promoter variants requires use of functional assays. To date, there are 13 *LDLR* promoter variants predicted to be causal (in revision[10]). Disappointingly, but not surprisingly, given they were not targeted, the exome sequencing data generated by the SureSelect Human All Exon (Agilent) assay, had negligible coverage of the gene promoter regions, which can lead to false negative conclusions. Further updates of the human exome capture assay should include coverage of the *LDLR* promoter sequences, which can cause autosomal dominant disease by altering gene regulation.

Exome sequencing

The SureSelect Human All Exons capture assay is a standard product, which proved to be efficient at detecting mutations within the *LDLR* and the *APOB* genes. In this sample, 78.9% of exome bases were covered at least 20 times, which is in line with the product description ~80%. For both *LDLR* and *APOB*, the majority of the coding sequence was covered more than the 16x threshold to achieve an estimated 99% chance of seeing a real variant (present in a heterozygous state) of at least 3 times, and overall 19 mutations, were found by high-throughput DNA sequencing, which had been missed by conventional methods in our research laboratory. This indicates increased sensitivity for NGS, which can be due both to the method used and to the reduced human intervention, and the highly automated protocol. However, as with many PCR-based methods, exome sequencing has some limitations when it comes to amplification of highly repetitive regions or sequences rich in GC content. A highly significant negative correlation between the G/C content and the exome depth was observed in the FH genes ($p=4.9\times10^{-14}$), as shown in the online supplementary figure S5. Specifically, only 58% of the *PCSK9* gene was covered more than 16x, producing unreliable results for variant calling in a significant proportion of the gene's coding region. As a result, the two novel non-synonymous *PCSK9* variants called by the exome sequencing were not confirmed by the capillary sequencing, suggesting a high rate of false positive calls when the coverage is poor. If a read depth threshold of 30x was considered to be required for complete certainty of variant calling, at which the sensitivity to detect heterozygous variants was shown to be 100%,[31] exons that would be insufficiently covered would also include exons 1, 14 and 18 of *LDLR*, exons 2 and 5 of *PCSK9*, and exons 2, 3 and 7 of *LDLRAP1*. Thus, although the quality of the produced data is good, validation of called variants in poorly covered regions is still necessary. Applying more stringent filters to the raw data increases the specificity of the calling. However, it may also lead to false negative results, since not all of the exome's regions are equally covered. Newer versions of the SureSelect assay show markedly improved coverage of exons that were previously poorly covered (unpublished data), so we can expect the sensitivity of exome sequencing to improve.

The Agilent SureSelect assay was efficient in capturing the exon–intron junctions, covering on average 80–100bps of the intronic regions. This was an advantage over our current screening protocol, and enabled us to detect a novel variant, the c.695-6_698del, which is partially positioned on the annealing site for the sequencing primer routinely used in our lab.

The methodology behind the ExomeDepth package[23] proved to be robust, and enabled us to use the exome data, which are composed of short sequence reads for exonic regions, to detect large gene rearrangements, which are known in the *LDLR* to be usually due to intronic Alu sequence misparing.[32 33] The method was shown to allow identification of heterozygous CNVs within the *LDLR* gene, which were missed by the currently used MLPA. However, in order to maximise the sensitivity and to minimise the noise created by technical variability between samples, CNV analysis by Exome Depth requires quality data of well-matched exomes (>6 samples), that is, sequenced under the exact same conditions.

The greater time efficiency of the exome sequencing is a significant advantage over the current screening methods. Although each called variant currently needs to be individually confirmed by Sanger sequencing before a mutation report can be prepared, analysing a number of patients in parallel in a short period of time is likely to be an efficient way forward for screening of heterogeneous FH patients. More importantly, limited use of manual checks and human intervention reduce the issues of possible human error. The cost efficiency of NGS is also increasing. The development of novel approaches of gene-targeted sequencing, using Illumina MiSeq platform, reduces not only the costs of sequencing itself but also the time spent on data analysis and computer storage requirements. The possibility of designing custom amplicons for each disease, recently offered by Illumina TruSeq Custom Amplicon or Agilent HaloPlex products, will also improve the capture of promoters and other regulatory regions, which could be omitted in whole exome eequencing.

Acknowledgements We thank Ebele Usifo for carrying out the functional analysis prediction for the novel *LDLR* variant. This study 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.

Contributors MF: writing of the manuscript, data analysis. VP: data analysis, CNVs calling. RAW: sample preparation. HAWN: patient selection and study samples provider. UK10K exome sequencing and data production. she: data analysis, project supervision, guarantor of the publication.

Funding SEH holds a chair funded by the British Heart Foundation, and SEH and RW are supported by the BHF (PG08/008). MF is funded by an MRC CASE award with Gen-Probe Life Sciences Ltd, and VP is partially funded by a MRC research grant (G1001158). HAWN is a NIHR senior investigator.

Competing interests None.

Ethics approval National Research Ethics Service, Cambridgeshire 2 Research Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

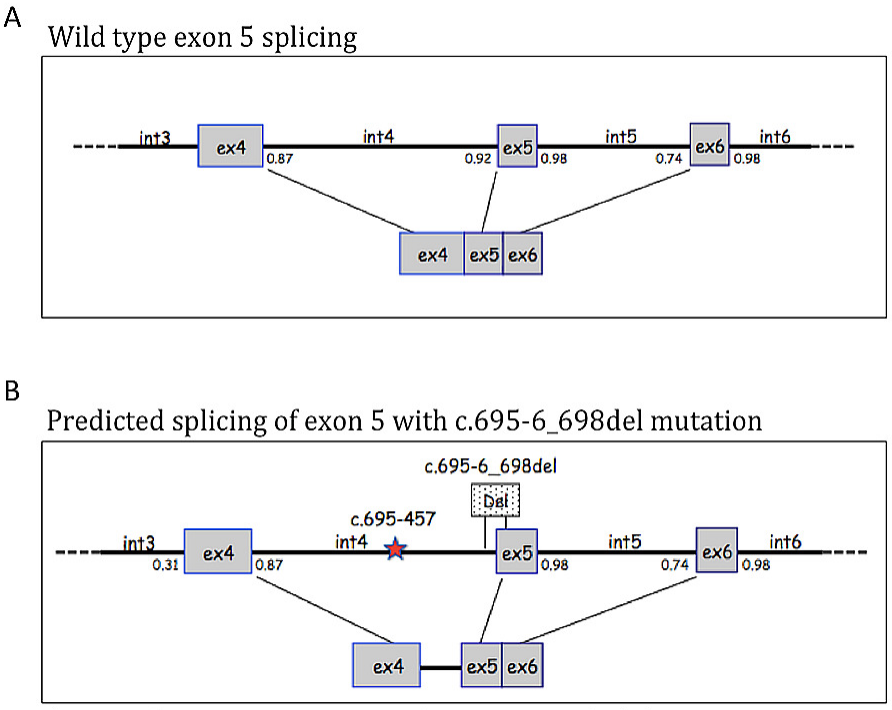
Data sharing statement The authors are happy to share the data with groups working on familial hypercholesterolaemia. All novel mutations were submitted to the publicly available FH mutation database (http://www.ucl.ac.uk/fh).

REFERENCES

1. **Marks D,** Thorogood M, Neil HA, Humphries SE. A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. *Atherosclerosis* 2003;**168**:1–14.
2. **Slack J.** Risks of ischaemic heart-disease in familial hyperlipoproteinaemic states. *Lancet* 1969;**2**:1380–2.
3. **Scientific Steering Committee on behalf of the Simon Broome Register Group.** Mortality in treated heterozygous familial hypercholesterolaemia: implications for clinical management. *Atherosclerosis* 1999;**142**:105–12.
4. **Oosterveer DM,** Versmissen J, Yazdanpanah M, Hamza TH, Sijbrands EJ. Differences in characteristics and risk of cardiovascular disease in familial hypercholesterolemia patients with and without tendon xanthomas: a systematic review and meta-analysis. *Atherosclerosis* 2009;**207**:311–17.
5. **Neil HA,** Huxley RR, Hawkins MM, Durrington PN, Betteridge DJ, Humphries SE. Comparison of the risk of fatal coronary heart disease in treated xanthomatous and non-xanthomatous heterozygous familial hypercholesterolaemia: a prospective registry study. *Atherosclerosis* 2003;**170**:73–8.
6. **Neil A,** Cooper J, Betteridge J, Capps N, McDowell I, Durrington P, Seed M, Humphries SE. Reductions in all-cause, cancer, and coronary mortality in statin-treated patients with heterozygous familial hypercholesterolaemia: a prospective registry study. *Eur Heart J* 2008;**29**:2625–33.
7. **Gill PJ,** Harnden A, Karpe F. Familial hypercholesterolaemia. *BMJ* 2012;**344**:e3228.
8. **Usifo E,** Leigh SEA, Whittall RA, Lench N, Taylor A, Yeates C, Orengo CA, Martin ACR, Humphries SE. Low density lipoprotein receptor gene familial hypercholesterolemia variant database: update and pathological assessment. *Ann Hum Genet* 2012;**76**:387–401.
9. **Humphries SE,** Cranston T, Allen M, Middleton-Price H, Fernandez MC, Senior V, Hawe E, Iversen A, Wray R, Crook MA, Wierzbicki AS. Mutational analysis in UK patients with a clinical diagnosis of familial hypercholesterolaemia: relationship with plasma lipid traits, heart disease risk and utility in relative tracing. *J Mol Med* 2006;**84**:203–14.
10. **Khamis APJ,** Lench N, Taylor A, Leigh S, Humphries SE. Analysis of four LDLR 5′UTR and promoter variants in patients with familial hypercholesterolaemia. 2012.
11. **Garcia CK,** Wilund K, Arca M, Zuliani G, Fellin R, Maioli M, Calandra S, Bertolini S, Cossu F, Grishin N, Barnes R, Cohen JC, Hobbs HH. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science* 2001;**292**:1394–8.
12. **Wierzbicki AS,** Humphries SE, Minhas R. Familial hypercholesterolaemia: summary of NICE guidance. *BMJ* 2008;**337**:a1095.
13. **Taylor A,** Patel K, Tsedeke J, Humphries SE, Norbury G. Mutation screening in patients for familial hypercholesterolaemia (ADH). *Clin Genet* 2010;**77**:97–9.
14. **Desai A,** Jere A. Next-generation sequencing: ready for the clinics? *Clin Genet* 2012;**81**:503–10.
15. **Graham CA,** McIlhatton BP, Kirk CW, Beattie ED, Lyttle K, Hart P, Neely RD, Young IS, Nicholls DP. Genetic screening protocol for familial hypercholesterolemia which includes splicing defects gives an improved mutation detection rate. *Atherosclerosis* 2005;**182**:331–40.
16. **Whittall RA,** Scartezini M, Li K, Hubbart C, Reiner Z, Abraha A, Neil HA, Dedoussis G, Humphries SE. Development of a high-resolution melting method for mutation detection in familial hypercholesterolaemia patients. *Ann Clin Biochem* 2010; **47**(Pt 1):44–55.
17. **Humphries SE,** Whittall RA, Hubbart CS, Maplebeck S, Cooper JA, Soutar AK, Naoumova R, Thompson GR, Seed M, Durrington PN, Miller JP, Betteridge DJ, Neil HA. Genetic causes of familial hypercholesterolaemia in patients in the UK: relation to plasma lipid levels and coronary heart disease risk. *J Med Genet* 2006;**43**:943–9.
18. **Taylor A,** Wang D, Patel K, Whittall R, Wood G, Farrer M, Neely RD, Fairgrieve S, Nair D, Barbir M, Jones JL, Egan S, Everdale R, Lolin Y, Hughes E, Cooper JA, Hadfield SG, Norbury G, Humphries SE. Mutation detection rate and spectrum in familial hypercholesterolaemia patients in the UK pilot cascade project. *Clin Genet* 2010;**77**:572–80.
19. **Neil HA,** Seagroatt V, Betteridge DJ, Cooper MP, Durrington PN, Miller JP, Seed M, Naoumova RP, Thompson GR, Huxley R, Humphries SE. Established and emerging coronary risk factors in patients with heterozygous familial hypercholesterolaemia. *Heart* 2004;**90**:1431–7.
20. **Mamotte CD,** van Bockxmeer FM. A robust strategy for screening and confirmation of familial defective apolipoprotein B-100. *Clin Chem* 1993;**39**:118–21.
21. **Taylor A,** Martin B, Wang D, Patel K, Humphries SE, Norbury G. Multiplex ligation-dependent probe amplification analysis to screen for deletions and duplications of the LDLR gene in patients with familial hypercholesterolaemia. *Clin Genet* 2009;**76**:69–75.
22. **Miller SA,** Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;**16**:1215.
23. **Plagnol V,** Curtis J, Epstein M, Mok K, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burn SO, Thrasher A, Kumararatne D, Doffinger R, Nejentsev S. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012.
24. **Usifo E,** Leigh SE, Whittall RA, Lench N, Taylor A, Yeats C, Orengo CA, Martin AC, Celli J, Humphries SE. Low-density lipoprotein receptor gene familial hypercholesterolemia variant database: update and pathological assessment. *Ann Hum Genet* 2012;**76**:387–401.
25. **Usifo EL,** SEA , Whittall RA, Lench N, Taylor A, Yeates C, Orengo CA Martin ACR, Humphries SE. Low density lipoprotein receptor gene familial hypercholesterolemia variant database: update and pathological assessment. *Ann Hum Genet 2012*.
26. **Barbagallo CM,** Emmanuele G, Cefalu AB, Fiore B, Noto D, Mazzarino MC, Pace A, Brogna A, Rizzo M, Corsini A, Notarbartolo A, Travali S, Averna MR. Autosomal recessive hypercholesterolemia in a Sicilian kindred harboring the 432insA mutation of the ARH gene. *Atherosclerosis* 2003;**166**:395–400.
27. **Innerarity TL,** Weisgraber KH, Arnold KS, Mahley RW, Krauss RM, Vega GL, Grundy SM. Familial defective apolipoprotein B-100: low density lipoproteins with abnormal receptor binding. *Proc Natl Acad Sci U S A* 1987;**84**:6919–23.
28. **Boren J,** Lee I, Zhu W, Arnold K, Taylor S, Innerarity TL. Identification of the low density lipoprotein receptor-binding site in apolipoprotein B100 and the modulation of its binding activity by the carboxyl terminus in familial defective apo-B100. *J Clin Investig* 1998;**101**:1084–93.
29. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
30. **Cooper GM,** Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**:628–40.
31. **Choi M,** Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009;**106**:19096–101.
32. **Lehrman MA,** Goldstein JL, Russell DW, Brown MS. Duplication of seven exons in LDL receptor gene caused by Alu-Alu recombination in a subject with familial hypercholesterolemia. *Cell* 1987;**48**:827–35.
33. **Lehrman MA,** Russell DW, Goldstein JL, Brown MS. Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J Biol Chem* 1987;**262**:3354–61.

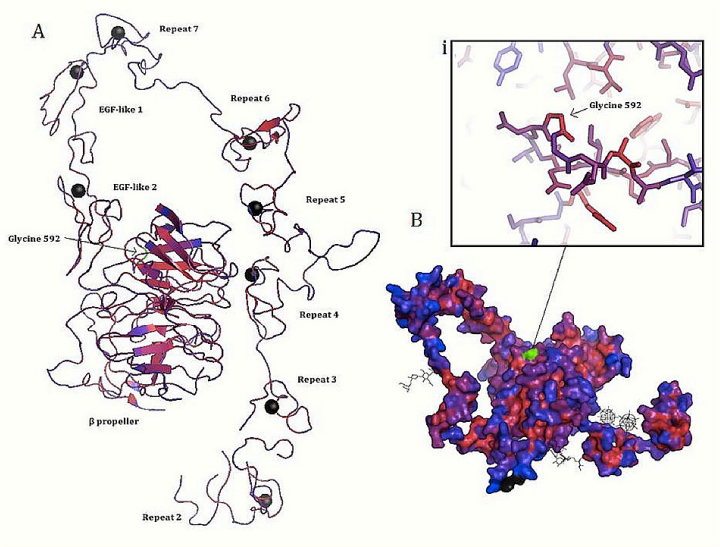
Supplementary Fig.S1:

In silico, splice site predictions using BDGP prediction tools (http://www.fruitfly.org/seq_tools/splice.html)[34]. A: Wild type *LDLR* scores for the acceptor/donor splice sites. B: *LDLR* mutation (c.695-6_698del) removes the exon 5 acceptor site and potentially activates an upstream cryptic site (c.695-457/8).



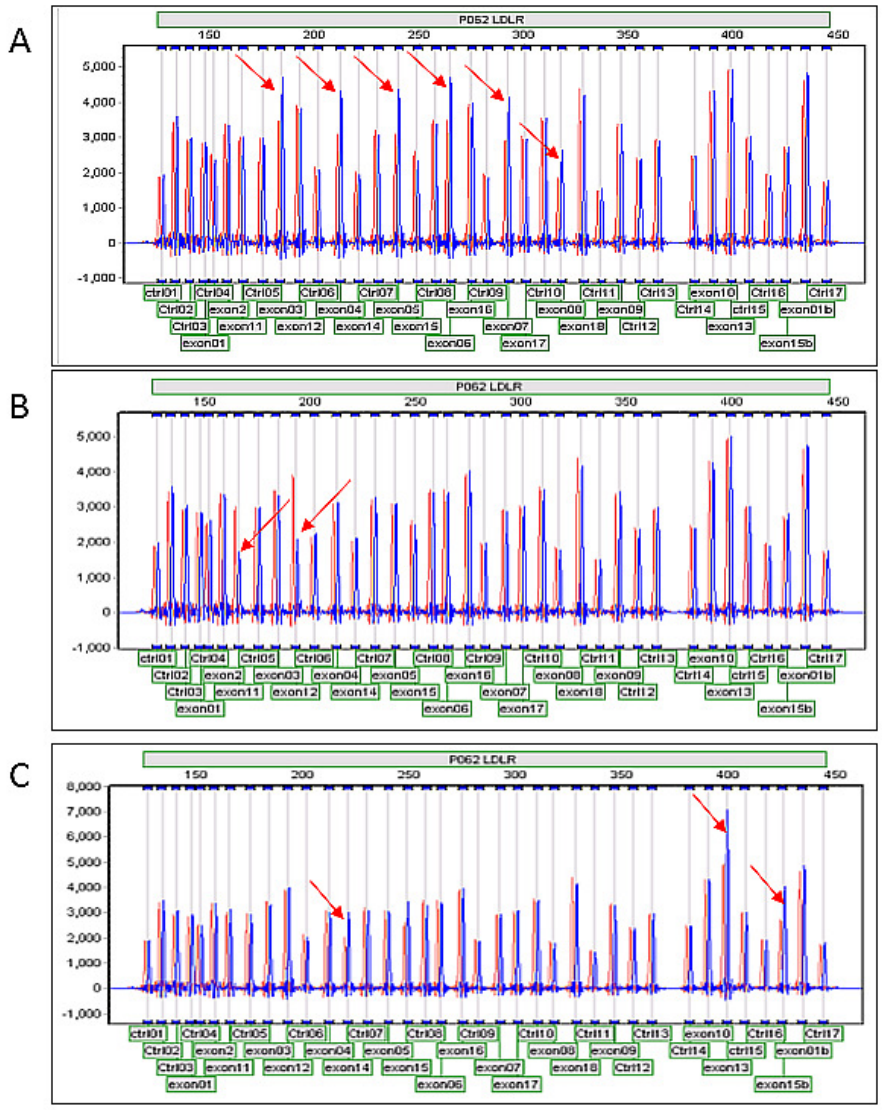
Supplementary Fig.S2:

Conservation score calculated over the entire available crystal structure of the LDL-R protein. Red shows high conservation, purple- moderate conservation, blue- poor conservation and black- no conservation. A. LDL-R protein displaying the calcium molecules (black spheres) and labeled regions including Glycine 592. B. Glycine 592 is on the surface of the protein and is highly conserved (see inset 'i').



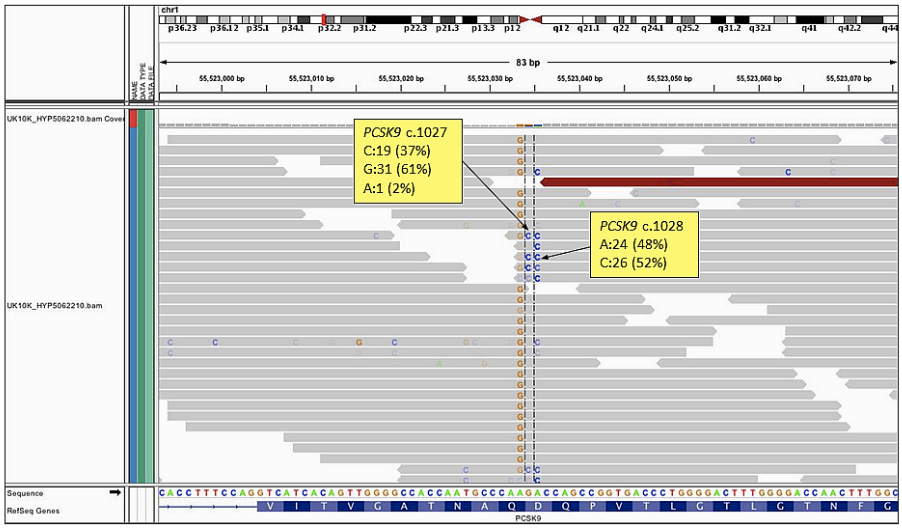
Supplementary Fig.S3:

MLPA results quantified by the fluorescence peak heights for the tested sample (blue) and normalised control (red). Red arrows mark peaks, in which the sample/control difference was significant. A: Heterozygous duplication of exons 3 to 8 in sample HYP5002209. B: Heterozygous deletion of exons 11 and 12 in sample HYP5062217. C: Heterozygous duplication of exons 13 to 15 in sample HYP5062219.



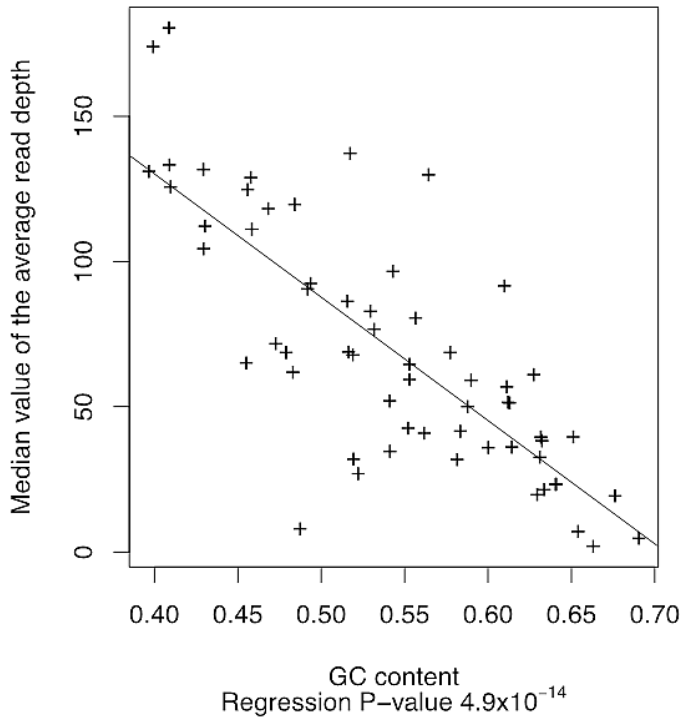
Supplementary Fig.S4:

Intergrative Genomic Viewer image of the coverage of *PCSK9* exon 7 region containing two false positive variants c.1027G>C and c.1028A>C. These artifacts were probably created during the amplification step in the sequence capture process.



Supplementary Fig.S5:

The negative correlation of the median read depth and the GC content for each targeted exon of the four FH genes (*LDLR*, *APOB*, *PCSK9* and *LDLRAP1*).



No. samples	Nucleotide change	Functional effect	Depth	Quality	PolyPhen	SIFT	Mutation Taster
1	c.148C>T	p.(R50W)	139	181	D	D	N
1	c.1199G>A	p(.R400H)	35	100	B	D	N
1	c.2938G>A	p.(A980T)	29	198	B	T	N
1	c.3931A>C	p.(K1311Q)	68	170	B	D	N

Table S1.

Summary of novel *APOB* variants, located outside of exons 26 and 29. 'Depth' refers to the depth coverage; 'Quality' values are Phred-like quality scores generated by SAMtools. D-damaging; B-benign; T-tolerated; N- polymorphism.