The use of new technology to improve genetic testing

Almomani, R.

Chapter **3**

Experiences with array-based sequence capture; toward clinical applications

Rowida Almomani, Jaap van der Heijden, Yavuz Ariyurek, Yuching Lai, Egbert Bakker, Michiel van Galen, Martijn H Breuning and Johan T den Dunnen

Eur J Hum Genet. 2011; 19: 50–55.

Abstract

Although sequencing of a human genome gradually becomes an option, zooming in on the region of interest remains attractive and cost saving. We performed array-based sequence capture using 385K Roche NimbleGen, Inc. arrays to zoom in on the protein-coding and immediate intron-flanking sequences of 112 genes, potentially involved in mental retardation and congenital malformation. Captured material was sequenced using Illumina technology. A data analysis pipeline was built that detects sequence variants, positions them in relation to the gene, checks for presence in databases (eg, db single-nucleotide polymorphism (SNP)) and predicts the potential consequences at the level of RNA splicing and protein translation. In the samples analyzed, all known variants were reliably detected, including pathogenic variants from control cases and SNPs derived from array experiments. Although overall coverage varied considerably, it was reproducible per region and facilitated the detection of large deletions and duplications (copy number variations), including a partial deletion in the *B3GALTL* gene from a patient sample. For ultimate diagnostic application, overall results need to be improved. Future arrays should contain probes from both DNA strands, and to obtain a more even coverage, one could add fewer probes from densely and more probes from sparsely covered regions.

Introduction

For many years, the amplification of target sequences by PCR, followed by Sanger sequencing, has been the gold standard for screening of variants in terms of both read length and accuracy of sequencing.[1] However, when it comes to conditions with highly heterogeneous etiology, a large number of different genes need to be screened for mutations. In such cases, gathering information becomes laborious, expensive and time-consuming. There are many examples of diseases that can be caused by mutations in many different genes, including mental retardation (MR),[2] Charcot–Marie–Tooth disease,[3] cardiomyopathy,[4] retinitis pigmentosa,[5] autism,[6] hearing loss[7] and congenital disorders of glycosylation.[8] Extensive resequencing of many disease-associated genes is required to explore, at the sequence and structural level, the genomic variation that might be involved in causing such diseases.

Several next-generation sequencing (NGS) platforms are now available and they have allowed the sequencing and analysis of large numbers of genes in one experiment,[9, 10, 11] and are able to

generate a massive amount of sequence data and have considerably reduced the cost of DNA sequencing.[12] However, although NGS platforms have enormously increased throughput and have permitted whole-genome sequencing, high cost still prevents routine whole human genome resequencing projects. Therefore, zooming in on the region of interest is an attractive option. In addition, it circumvents the problem of identifying variants in genes for which the analyses were not intended (with associated ethical problems).

Microarray-based genomic selection combined with massively parallel high-throughput sequencing is the method of choice to analyze large numbers of genes in a more comprehensive and cost-effective manner.[13, 14, 15] We have used custom high-density microarrays (Roche NimbleGen, Inc., Madison, WI, USA) for the enrichment of 112 distinct genes potentially involved in MR and congenital malformation, followed by sequencing on the Illumina Genome Analyzer I platform (Illumina, San Diego, CA, USA).

The first aim of our study was to apply and validate the array-based enrichment method as an efficient and convenient strategy to capture any desired portion of the human genome. The second aim was to accelerate the detection of sequence and copy number variations (CNV) in the selected candidate genes with lower costs, especially for the genes that are potentially involved in MR.

Materials and methods

Sample selection and validation

Six DNA samples were used in this study, including two controls containing known pathogenic variants. Sample S-2 contains a known *MECP2* (OMIM 300005) pathogenic point mutation (c.538C>T); the second sample, patient S-6, carries a large deletion spanning exons 8–15 in one allele and a splice site mutation (c.660+1G) at the other allele of the *B3GALTL* (OMIM 610308) gene.

The other four DNA samples were from patients with MR with an unknown cause. Single-nucleotide polymorphism (SNP) array data were available for two samples: S-7 with 250K Nsp Affymetrix and S-5 with 317K Illumina data. We used these data to validate the sequences

obtained after capture-array and Illumina sequencing. Causative large deletions and duplications had been previously excluded by SNP array testing in S-3, S-5, S-7 and S-8.

Exon array design

Microarrays with 385K probe capacity (Roche NimbleGen, Inc.) were used to capture all exons, the splice site and the immediately adjacent intron sequence of 112 human genes. On the basis of searches in OMIM and literature, we selected 112 human genes known to cause MR, either as part of a known syndrome or in isolation (Supplementary Table 1). Primary sequence data from all exons were extracted from NCBI's genome (Build 36). Microarrays were designed by Roche NimbleGen, Inc. with long oligonucleotide probes (54–99 nucleotides) that span each target region, overlapped and shifted on an average of seven bases.[13] The oligonucleotides were designed to achieve isothermal hybridization across the arrays capturing one strand only. All highly repetitive regions were excluded from the probe selection in order to avoid nonspecific capturing of genomic regions. Using all criteria listed, for 2% of the target sequences, no capture probe could be designed (note that, theoretically, these sequences can be covered partly through capture from directly flanking unique sequences). Four of the arrays were reused at least twice.

Genomic DNA library preparation and target capture

The methods used for target capture, enrichments and elution followed previously described protocols with slight modifications (Roche NimbleGen, Inc.).[16] Genomic DNA (20–10 μg) was fragmented using a nebulizer or Bioruptor according to instructions from the manufacturer to yield fragments from 250–1000 bp (nebulization) or 250–600 bp (Bioruptor). Adapter oligonucleotides from Illumina (single reads) were ligated to the ends. After the ligation was completed, successful adapter ligation was confirmed by PCR. The DNA-adapter ligated fragments were then hybridized to the sequence capture microarray for 65 h. After hybridization and washing, the DNA fragments bound to the array were eluted, using 300 μl of the elution buffer (Qiagen, Valencia, CA, USA) on each array. A gasket (Agilent) was applied and placed on the thermal elution device (homemade) for 20 min at 95°C. We repeated this process once by adding 200 μl of elution buffer (Qiagen). DNA from each eluted sample was enriched by 18-cycle PCR using a high-fidelity polymerase and a single primer pair corresponding to the Illumina adapters ligated earlier.

Check enrichments by qPCR

To verify successful hybridization capture, we performed qPCR (quantitative PCR) on DNA samples (S-2, S-3, S-5, S-7, S-6 and S-8) before and after array enrichment. The primers amplified five loci from *MBL2*, *DMD* and *BRCA1* (100 bp) as negative controls (no capture probes on the array) and four loci from *MECP2*, *CREBBP* and *NSD1* genes as positive controls (capture probes on the array) (Supplementary Table 2). All primers for qPCR were designed using Primer 3 (http://frodo.wi.mit.edu/).

The qPCR assays were performed in triplicate in the Lightcycler using 384-well plates (Roche NimbleGen, Inc.) in 10 μl total volume: 5 μl of 2 × SYBR Green master Rox (Roche NimbleGen, Inc.), 0.25 μl of each primer (10 pmol/μl), 2 μl of DNA template and 2.5 μl of ultrapure water. The thermo-cycling protocol was carried out as follows: 10 min at 95°C, 45 cycles of 10 s at 95°C, 30 s at 60°C, 20 s at 72°C and 5 min at 72°C, followed by melting curve analysis in order to determine the specific and nonspecific amplified products and other artifacts that might interfere with CP values. To calculate the relative fold enrichment of the targeted regions, we compared amplification of the positive *versus* negative controls. The relative fold enrichment, R, was calculated using the values of ΔCP (ie, the difference between average CP of non-captured and average CP of captured samples) according to $R=E^N$, where E is the efficiency of the qPCR assay for a particular amplicon and N=ΔCP (crossing point).

DNA Sequencing

The eluted enriched DNA fragments were sequenced using the Illumina GAI platform at the Leiden Genome Technology Center (LGTC). Single-end sequencing of 36 or 50 nucleotides was performed following the instructions of the manufacturer.

Reads mapping and data analysis

Sequence read mapping was carried out by ELAND and ELAND-extended programs, which were a part of the Illumina GAI data analysis package. Only reads of high-quality scores were mapped to the human reference genome (NCBI, BUILD 36.2), allowing up to two mismatches. We created different Perl scripts to extract and process data from the ELAND files. Coverage was calculated at the target level (gene–exons), the nucleotide level and at the per probe region.

SNP calling was performed by searching for nucleotides discordant with the reference genome with a base call quality score of 30 (99.9% base call accuracy), a read depth of 8 or greater and the variant allele larger than 30% of the total coverage. Thereafter, all variants were checked for their presence in known databases, for example, dbSNP. Perl scripts were designed to predict the potential consequences at the level of RNA splicing and protein translation on the basis of Ensemble v.51. Furthermore, we designed a Perl script to facilitate detection of small deletions/insertions (up to three nucleotides). All Perl scripts are available on request.

Sanger sequencing

A total of 21 variants detected by Illumina GAI analyzer were selected and confirmed by Sanger sequencing using the standard Sanger sequencing protocol at the Leiden Genome Technology Center (LGTC). The primer sequences (with M13 tail) used are shown in Supplementary Table 3.

Results

The methodology used starts with fragmentation of the genomic DNA. Linker and primer addition can then be performed either before or after array-capture target enrichment. To facilitate limited amplification of the expected low-yield array elution, we decided to perform full Illumina sample preparation before array capture. Initially, experiments were conducted using 20 μg genomic DNA, later we reduced this to 10 μg. We used qPCR, comparing targeted (four positive controls) and non-targeted regions (five negative controls), to check successful array enrichment and to estimate the fold enrichment obtained (see Supplementary Tables 4 and 5 for examples). As enrichment varies significantly from locus to locus, we tested multiple loci to obtain an accurate estimate. Samples in which qPCR did not indicate clear enrichment (>100 ×) were discarded. The ultimate enrichments achieved varied from experiment to experiment with a tendency to increase over time, indicating that lab experience is an important aspect of the array capture technology. As the fold enrichments determined by qPCR correlate positively with

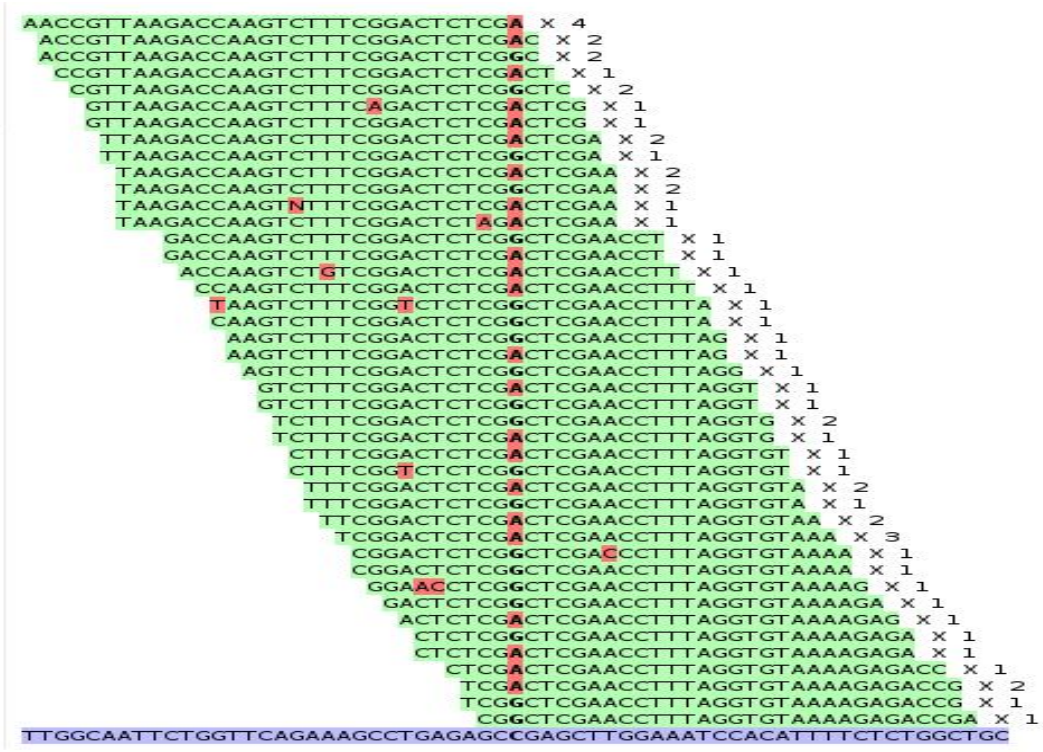

Figure 1 Detection of sequence variants. A total of 32 nucleotide NGS reads (top, sequence mismatches in red) aligned with the genomic reference sequence (bottom). The center of the alignment shows a variant present in the heterozygous state. ' × *n*' behind the read indicates how many identical reads were obtained.

the average sequence depth obtained, we conclude that qPCR provides an effective and cost-saving check for successful enrichment (examples are listed in Supplementary Tables 4 and 5).

Sequence data

The custom arrays used contained 112 different human genes that are known to be or potentially involved in MR and congenital malformation. Samples were run on one channel of the Illumina GAI. For sequence analysis, we used only those QC-filtered reads that map back uniquely to the reference sequence (M0) or with one or two mismatches (M1, M2) (Figure 1). Using these settings, 85–92% of the targeted nucleotides were covered by at least eight reads (Table 1) and 94–98% by at least one read (note that for 2% of the targeted sequences, no probe could be designed, see M&M). Effectively, this means that for 78% of the targeted sequences on the array, coverage was sufficient (>20 ×) to detect any variants that were present.

Table 1 Sequence summary results of the different array-capture experiments performed
Abbreviations: F, female; M, male; MM# reads, number of reads with # mismatches to the reference sequence; QC, quality control.

Sample ID, sex	Total reads × 10^3	Reads passing QC filter × 10^3	Total number of reads mapped × 10^3	MM0 reads × 10^3	MM1 reads × 10^3	MM2 reads × 10^3	Coverage per nucleotide	% of Nucleotides were covered ≥8 times	% of nucleotides were covered 0 times	Read length	Array reused
S-2, F	6.744	4.804	2.428	1.359	691	378	138	87.11	6.22	50	No
S-3, M	7.305	5.354	2.176	1.225	618	333	100	90.71	4.49	50	No
S-5, M	10.43	7.237	5.576	4.935	499	142	120	92.42	2.09	32	No
S-7, M	15.771	6.112	4.719	3.885	638	196	100	91.13	2.7	32	No
S-6, M	12.154	6.575	6.575	5.914	486	174	99	99.24	7.08	32	Yes, 2nd time
S-8, F	11.077	3.531	3.531	2.301	736	485	44	85.38	4.43	49	Yes, 3rd time

Two of the samples had been previously analyzed using SNP arrays. The region selected using the capture array included 67 different SNPs that had been present on the SNP arrays. We observed a perfect agreement (100%) between array-based SNP calls and those obtained using NGS (67/67 variants) (Supplementary Table 6).

To determine our ability to detect pathogenic mutations, we included one sample from a female patient (S-2) harboring a dominant pathogenic point mutation in the *MECP2* gene, (c.538C>T) on the X chromosome. Our results clearly detected the change in the heterozygous state (Supplementary Table 7). Similarly, we detected a homozygous change in the *B3GALTL* gene in a Peter's Plus patient (c.660+1G>A, Supplementary Table 7, see below).

We next selected 21 variants detected in samples S-2, S-3, S-5, S-7 and S-8 and checked these by traditional Sanger sequencing. We were able to confirm 21 of the 21 variants, including their status being homozygous or heterozygous (Supplementary Table 7). The analysis of the variants found in all 112 genes of the patients did not reveal a clear cause of their MR Supplementary Table 8 and 9.

CNV

Changes that cannot be easily detected using the sequence itself include deletions and duplications (CNVs). However, such variants can be expected to yield quantitative changes in coverage. To determine whether overall coverage can be used to detect quantitative changes, we first analyzed the 39 genes located on the X chromosome. Indeed, when coverage was normalized using autosomal genes (Figure 2a), samples from females showed a clearly higher X-chromosome coverage compared with male samples (Figure 2b). Furthermore, as expected, the gene on the Y chromosome (*NLGN4Y*) gave no coverage in the female sample (Figure 2b). To determine the sensitivity of our method for detecting smaller CNVs, we carefully analyzed a sample from a compound heterozygous patient (S-6) carrying a partial deletion (exons 8–15) and a splice site mutation (c.660+1G>A, intron 8) in the *B3GALTL* gene. The splice site mutation was evident as no wild-type sequence was present. The presence of a deletion emerged as, compared with other samples, we observed a significantly lower average coverage for the *B3GALTL* gene (53 × *versus* 155 ×, 150 ×, 140 ×) (Figure 2c). In addition, although the splice site mutation in exon 8 was detected in the 'homozygous' state (similar to all nine variants downstream), we observed variants in the first exons (1–7) also in heterozygous state (Supplementary Table 10). These data show that not only have we obtained an excellent specificity of the capture process but we have also been able to distinguish between male and female samples.

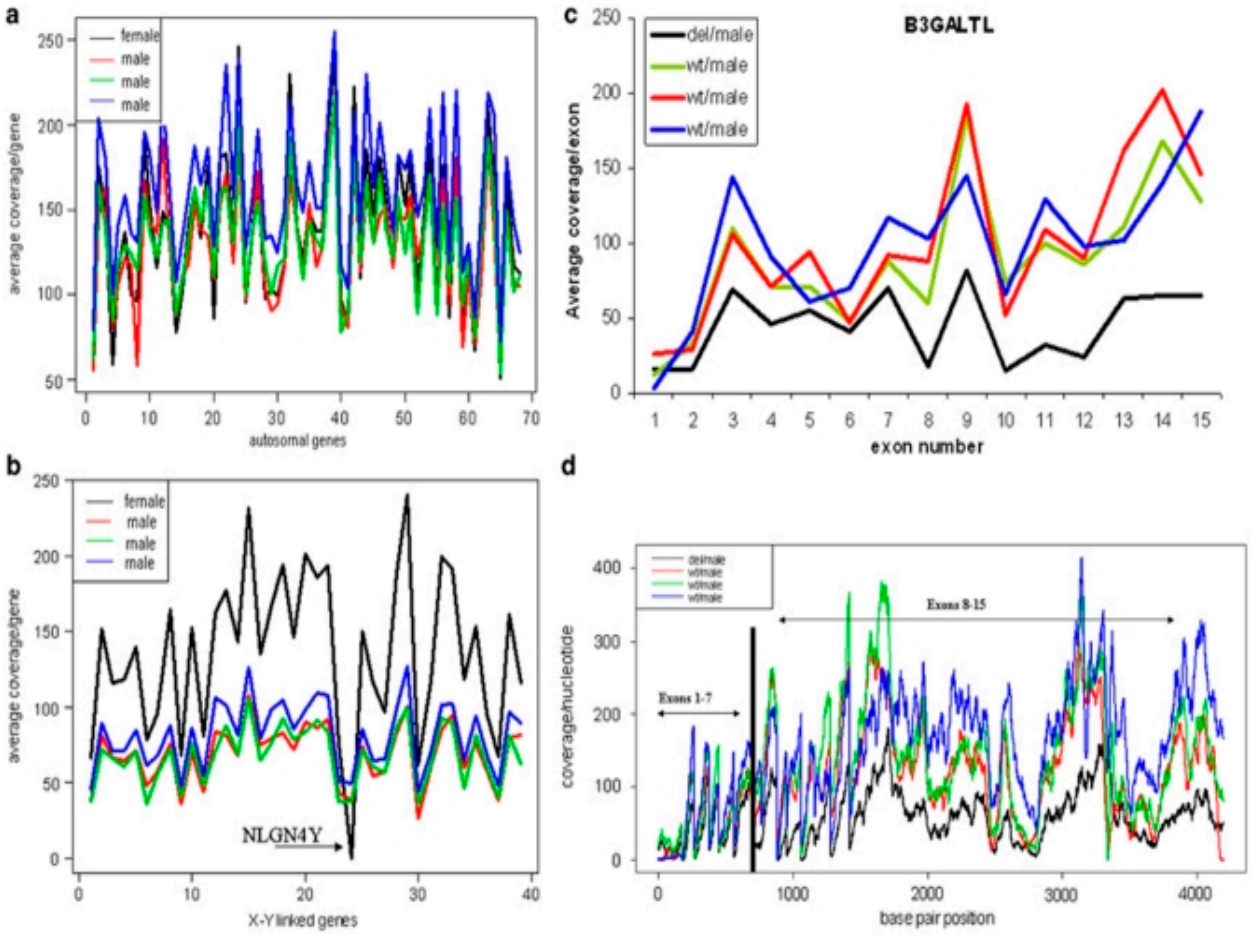

Figure 2 Average coverage obtained for different genes in four different samples. (**a**) Shows average coverage of 69 autosomal genes from four different samples. (**b**) Shows average coverage of 39 genes located on X and one gene (*NLGN4Y*) located on the Y chromosome; a female sample exhibited an absence of hybridization on the captured array, with no coverage in the regions corresponding to the *NLGN4Y*. The female sample shows a higher average coverage per gene for all genes located on X-chromosome compared with male samples. (**c**) Lower average coverage of *B3GALTL* gene in a male patient sample with a known large deletion compared with three wild-type male samples. (**d**) Coverage per nucleotide/position for the whole *B3GALTL* gene: the patient sample shows lower coverage for the second half (exons 8–15) compared with wild type samples. del=deletion, wt=wild type.

Discussion

Array-based genomic selection offers several advantages for large-scale targeted DNA isolation over other approaches such as PCR-based methods (long-range PCR or multiplexed short PCR),[17, 18, 19] selector technology[20, 21] and BACs technology.[22] PCR-based methods become

laborious, time-consuming and costly if hundreds to thousands of regions (exons) need to be amplified, especially if all the sequences are required. Furthermore, when PCRs are multiplexed, it becomes difficult to check successful amplification per fragment, the chance of obtaining artifacts increases and equimolar loading before sequencing becomes very difficult. New approaches for massive individual PCR have been introduced recently[23] but experiences with these are still limiting. Selector technology[20, 21] seems attractive but it largely depends on proper in-house probe design, and experience thus far is very limited. Successful genomic selection using BACs has been demonstrated but has several limitations. As a BAC is the unit of selection, multiple BACs are required to isolate discontinuous regions of interest.

In this study, we have tested array-based sequence capture to determine the sequence of 112 genes potentially involved in MR. We show that array-based sequence capture technology is an efficient, quick and reliable method for the parallel sequencing of a range of genes of interest. Known variants (array-based calls) for 67 SNPs matched perfectly with those obtained using NGS Supplementary Table 6. Two positive controls with known pathogenic changes in the *MECP2* gene (sample S-2) and *B3GALTL* gene (sample S-6) were readily detected. In addition, 21/21 selected variants found in the five samples analyzed could be confirmed using Sanger sequencing (Supplementary Table 7). Sequence coverage of the nucleotide of interest is critical for reliably detecting sequence changes. If coverage is too low, both false positives (caused by sequence errors) and false negatives (if only one allele from a heterozygous sample is observed) will occur.

The coverage we obtained differs significantly not only between targeted genomic regions (genes) but also between different samples (Supplementary Table 1, Figure 2a). As the overall methodology is rather complex, particularly the collection of the hybridized array-enriched DNA sequences, the difference between samples is most probably influenced by technical factors such as variations in hybridization, washing conditions and potential reuse of the capture array. Furthermore, coverage is influenced by array design, including probe sequence (melting temperature, GC content), probe density and spacing (Supplementary Table 1). Our data show that AT-rich regions (>55%), regions with an overall low probe density (<3) and small exons (on average 90 bp) yield a low coverage, which also varies significantly between experiments. For a second-generation capture array, the results obtained could be used to change the probe density,

that is, decreased in well-covered and increased in low-covered regions. Our data show that longer reads (50 bp) improve accuracy and selectivity of read mapping to the reference genome, which influenced the SNP calling by having less false positives and slightly better coverage.

As CNVs (deletions/duplications) are a significant cause in the etiology of MR,[24] we tested the feasibility of detecting large CNVs using array capture and NGS. Our results indicate that, if coverage is sufficiently high, array capture can also be used to detect such quantitative changes. Our array contained one gene from the Y chromosome that gave no coverage in females (Figure 2b), whereas the 39 X-linked genes when compared with the 69 autosomal genes yielded overall 50% lower coverage in male samples (Figure 2b). Another example derives from a sample containing a partial *B3GALTL* gene deletion on one allele (exons 8–15) and a splice site mutation on the other allele (c.660+1G>A). Although coverage over the entire gene seems reduced (experimental variation/coincidence), coverage for the second half of the gene clearly drops below that of normal (Figure 2d). An algorithm for detecting local deviations from the average coverage is currently under development.

Regarding probe design (performed by Roche NimbleGen, Inc.), it should be noted that all array probes are from one strand (coding DNA strand) and thus DNA molecules from only the non-coding strand are captured. This has several consequences. First, the sequence obtained is from one strand only, whereas for diagnostic applications, quality assurance requires that sequences be obtained in forward and reverse orientation. Sequencing this one strand in both directions is partly fooling oneself. Second, we observed that the sequences obtained relative to the array probes extend in a 5′ but not in a 3′ direction. The most probable cause for the latter is steric hindrance during array hybridization, preventing non-hybridizing tails at the surface side of the array. When capture probes are attached with their 3′ ends, this has consequences for probe design at the edges of the targeted regions; on the 5′ side, coverage will be significantly better than on the 3′ side. Both effects could be overcome simply by reversing the probe sequence of every other nucleotide on the array. Theoretically, this would also mean that the overall yield of enriched DNA would double, as both strands from the sample will be captured.

To save costs, we have reused the arrays up to three times by hybridizing different samples. The danger of this approach is of course contamination, if hybridized DNA from a previous

experiment is not eluted completely. Indeed, in some experiments, we observed low-level contamination, for example, through heterozygous calls from X-chromosome sequences in male samples. It should be noted, however that cross-contamination can be easily controlled when samples containing differently tagged linkers are used in subsequent experiments.

Using the current design, low coverage was obtained mainly at the edges of the regions targeted, especially the 3′ side (see above), that is, direct gene flanking or intronic regions. Although coverage varied widely, 78% of all regions targeted and present on the array were covered effectively by the sequence obtained. Note that there is a clear correlation between fragment size of the genomic DNA used and the coverage, the larger the fragment size used the lower the target coverage achieved, as more flanking DNA is captured. Especially for array-based capture, because of the steric hindrance described, this effect will be significant near the array-attached end of a probe-targeted region. Assuming that second-generation capture arrays will be more effective (ie, complete and with even coverage) and sequence power will improve further, it should soon be possible to sequence-tag, mix and simultaneously analyze different samples in one experiment, giving a significant cost reduction.

Recently in-solution capture was presented as an alternative to array-based capture.[25] Besides advantages of simplicity, a reduced workload and a potential for automation, when attempted, in-solution capture will not show the effect of steric hindrance we observed. However, capturing both strands would be complicated by the fact that capture probes will hybridize with each other. Initial experiences in our lab with in-solution capture were successful and for future projects we will change to this approach.

Overall, we conclude that array-based sequence capture followed by NGS offers a versatile tool for successfully selecting sequences of interest from a total human genome. The approach will be especially helpful in speeding up the identification of the pathogenic mutation(s) in diseases in which the genomic region to be scanned is large. Our results indicate that the methodology can still be improved, in particular, with respect to probe design, obtaining a more even coverage of the targeted regions. On the basis of initial experiences and publications, we expect that array capture will be quickly replaced by in-solution capture. Ultimately, the cost of this approach is

determined by the minimal coverage, which in turn determines the sensitivity required for the detection of potential sequence variants.

Acknowledgments

We thank the Leiden Genome Technology Center (LGTC), in particular Sophie Greve-Onderwater, Matthew Hestand and Rolf Vossen, for their expert technical assistance; Antoinette Gijsbers for sharing the SNP data; and Kamlesh Madan for critical reading of the paper. The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under Grant agreements 223026 (NMD-chip) and 223143 (the TechGene).

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–5467.
2. Chelly J, Khelfaoui M, Francis F, Chérif B, Bienvenu T. Genetics and pathophysiology of mental retardation. Eur J Hum Genet. 2006;14:701–713.
3. Szigeti K, Lupski JR. Charcot-Marie-Tooth disease. Eur J Hum Genet. 2009;17:703–710.
4. Paul M, Zumhagen S, Stallmeyer B, Koopmann M, Spieker T, Schulze-Bahr E. Genes causing inherited forms of cardiomyopathies. A current compendium. Herz. 2009;34:98–109.
5. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet. 2006;368:1795–1809.
6. Muhle R, Trentacoste SV, Rapin I. The genetics of autism. Pediatrics. 2004;113:472–486.
7. Hilgert N, Smith RJ, Van Camp G. Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics. Mutat Res. 2009;681:189–196.
8. Freeze H. Genetic defects in the human glycome. Nat Rev Genet. 2006;7:537–551.
9. Bonetta L. Genome sequencing in the fast lane. Nat Methods. 2006;3:141–147.
10. von Bubnoff A. Next-generation sequencing: the race is on. Cell. 2008;132:721–723.
11. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5:16–18.
12. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–1145.
13. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 2007;4:903–905.
14. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods. 2007;4:907–909.
15. Hodges E, Xuan Z, Balija V, et al. Genome-wide *in situ* exon capture for selective resequencing. Nat Genet. 2007;39:1522–1527.
16. Roche NibleGen NimbleGen services user's guides: sequence capture service . http://www.nimblegen.com/products/lit/SeqCap_UsersGuide_Service_v3p0.pdf .
17. Edwards MC, Gibbs RA. Multiplex PCR: advantages, development, and applications. PCR Methods Appl. 1994;3:S65–S75
18. Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal. 2002;16:47–51.
19. Cutler DJ, Zwick ME, Carrasquillo MM, et al. High-throughput variation detection and genotyping using microarrays. Genome Res. 2001;11:1913–1925.
20. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res. 2005;33:71.
21. Dahl F, Stenberg J, Fredriksson S, et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci USA. 2007;104:9387–9392.
22. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. Nat Methods. 2005;2:63–69.

23. Tewhey R, Warner JB, Nakano M, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol. 2009;27:1025–1031.

24. Shaw-Smith C, Redon R, Rickman L, et al. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. J Med Genet. 2004;41:241–248.

25. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–189.

Supplementary data:

Supplementary Table 1 Average coverage for all selected genes for six different samples. S-2, S-8, are female samples while the rest are male samples. All genes with bold have low GC or /and low probe density. Columns in bold represent data from re-used arrays.

Gene Name	Chromosome location	**S_6**	S_2	S_3	S_5	S_7	**S_8**
ACSL4	X	**37.58**	66.24	37.68	45.04	55.24	**32.2**
AFF2	X	**81.6**	152.41	72.51	89.61	92.34	**57.94**
AGTR2	X	**66.93**	115.49	66.91	71.58	96.6	**48.06**
ALG1	16	**55.84**	61.43	62.68	79.63	46.89	**21.58**
ALG12	22	**147.22**	176.4	164.69	203.65	85.58	**41.45**
ALG2	9	**163.81**	147.3	139.8	180.21	186.49	**54.07**
ALG6	1	**80.79**	59.33	84.49	99.93	129.98	**37.87**
ALG8	11	**105.88**	113.45	118.26	140.59	136.9	**48.35**
ALG9	11	**121.78**	137.43	129.53	158.8	143.43	**52.15**
AMMECR1	X	**64.75**	118.66	60.42	70.6	83.18	**49.71**
ARHGEF6	X	**71.53**	139.8	71.27	84.35	82.22	**56.31**
ARX	X	**47.8**	78.39	35.73	61.96	20.88	**23.93**
ASPM	1	**117.55**	100.77	124.26	136.38	180.53	**52.79**
B3GALTL	13	**58.33**	96.4	114.61	130.66	155.11	**40.4**
B4GALT1	9	**167.08**	193.19	159.43	195.85	144.86	**52.82**
BBS1	11	**142.78**	153.75	138.72	183.65	107.54	**41.16**
BBS10	12	**136.27**	116.27	121.91	148.89	186.85	**49.54**
BBS12	4	**191.59**	149.38	141.62	216.64	235.1	**71.96**
BBS2	16	**135.96**	139.07	143.86	167.26	155.66	**52.06**
BBS7	4	**87.54**	78.35	88.24	107.63	132.32	**40.22**
BRWD3	X	**57.78**	95.96	56.43	68.09	72.16	**47.02**
CA2	8	**111**	102.69	108.39	133.41	126	**42.57**
CC2D1A	19	**117.69**	140.17	130.99	158.96	66.25	**29.97**

CDK5RAP2	9	**153.43**	162.19	163.23	187.89	146.64	**58.83**
CDKL5	X	**75.99**	165.03	73.1	88.4	70.16	**54.41**
CENPJ	13	**138.59**	133.45	141.94	164.93	176.59	**54.91**
COG7	16	**135.39**	176.24	163.72	186.49	111.66	**51.73**
CRBN	3	**108.94**	86.16	109.99	128.55	152.04	**38.71**
CREBBP	16	**153.08**	180.96	159.75	182.66	109.2	**48.82**
CUL4B	X	**36.41**	68.41	40.24	47.46	49.6	**34.64**
DHCR7	11	**171.56**	183.15	163.11	234.75	106.68	**50.25**
DLG3	X	**69.9**	153.26	75.57	86.07	57.36	**47.41**
DNMT3B	20	**118.41**	129.94	130.4	156.44	95.96	**36.4**
DPAGT1	11	**193.16**	245.7	207.64	239.34	156.49	**63.91**
DPM1	20	**97.19**	95.5	99.55	127.71	125.07	**45.49**
DYRK1A	21	**138.62**	140.82	139.63	158.83	171.06	**55.53**
EP300	22	**172.71**	171.73	155.09	197	138.73	**49.63**
ERCC8	5	**103.87**	97.76	117.7	132.83	146.93	**48**
FGFR3	4	**91.04**	104.19	100.77	134.5	53.29	**25.78**
FKTN	9	**94.39**	99.05	116.64	125.17	128.33	**43.87**
FMR1	X	**43.64**	80.63	48.33	54.48	68.83	**42.42**
FOXP2	7	**116.96**	118.59	120.86	150.1	134.51	**46.95**
FTSJ1	X	**84.3**	162.65	71.18	105.87	67.19	**50.11**
GDI1	X	**81.34**	177.75	88.1	102.05	52.92	**47.42**
GLI3	7	**170.51**	229.96	193.16	214.24	119.51	**57.33**
GRIA3	X	**68.43**	143.32	68.85	81.17	76.74	**56.3**
GRIK2	6	**141.74**	140.26	138.97	164.42	157.43	**59.8**
HRAS	11	**110.45**	119.66	109.32	148.37	50.77	**26.55**
HSD17B10	X	**108.3**	232.65	105.43	126.58	76.49	**58.98**
IL1RAPL1	X	**75.14**	135.02	64.74	79.09	81.97	**56.15**
JAG1	20	**153.84**	149.33	143.06	177.83	134.17	**48.77**

JARID1C	X	**78.79**	166.5	75.9	98.34	55.94	**40.82**
KRAS	12	**37.5**	27.35	38.21	55.47	74.39	**14.28**
L1CAM	X	**83.47**	194.58	92.76	104.82	46.23	**44.07**
MAOA	X	**72.42**	146.33	78.53	85.28	89.76	**57.2**
MCPH1	8	**116.95**	137.58	132.3	150.93	134.11	**50.26**
MECP2	X	**90.3**	202.17	83.87	98.16	70.25	**55.29**
MED12	X	**87.01**	185.32	90.93	109.51	70.96	**53.62**
MGAT2	14	**128.29**	137.63	126.84	151.21	129.3	**43.37**
MPDU1	17	**191.91**	201.65	178.4	210.59	143.1	**54.87**
MPI	15	**208.85**	238.66	216.49	254.72	147.26	**62.87**
MYCN	2	**94.44**	97.38	78.66	117.71	65.2	**25.33**
NF1	17	**81.22**	82.78	89.34	103.68	103.15	**35.33**
NLGN3	X	**92.27**	193.94	83.89	107.85	60.81	**48.03**
NLGN4X	X	**43.01**	76.28	37.59	50.9	54.92	**28.06**
NLGN4Y	Y	**38.17**	0	37.83	49.38	59.67	**0**
NSD1	5	**184.55**	222.02	187.13	208.31	179.62	**72.56**
NUFIP1	13	**7.36**	7.19	11.08	18.4	21.56	**3.08**
OPHN1	X	**74.32**	150.2	72.16	85.18	77.89	**54.7**
PAFAH1B1	17	**113.96**	109.42	118.84	136.76	160.18	**42.43**
PAFAH1B3	19	**162.95**	185.67	166.94	229.95	92.43	**43.41**
PAK3	X	**54.5**	115.72	60.31	64.4	64.17	**46.98**
PHF6	X	**57.49**	96.99	57.08	65.4	87.77	**47.64**
PHF8	X	**83.55**	188.65	87.66	98.96	67.6	**57.96**
PMM2	16	**130.66**	146.18	130.15	162.48	104.29	**40.29**
POMT1	9	**146.51**	180.39	171.95	201.25	108.67	**47.81**
PQBP1	X	**100.71**	240.55	99.96	127.36	56.04	**52.63**
PRPS1	X	**26.32**	61.85	35.36	44.24	38.85	**23.34**
PRSS12	4	**152.22**	154.55	139.87	171.21	143.31	**52.15**

PTPN11	12	**10.42**	9.97	13.07	20.86	18.3	**3.7**
RAB3GAP2	1	**116**	116.02	120.94	136.63	151.15	**47.11**
RAF1	3	**147.08**	180.38	158.42	182.34	130.38	**54.88**
RAI1	17	**144.02**	152.31	124.16	173.36	70.74	**31.93**
REST	4	**158.76**	178.87	150.5	185	161.32	**55.18**
RNF135	17	**121.02**	101.12	88.77	145.5	117.24	**43.3**
RPS6KA3	X	**60.25**	117.99	60.92	70.45	83.05	**47.97**
SATB2	2	**139.83**	150.23	142.16	164.9	156.18	**54.62**
SCN8A	12	**170.77**	187.94	171.23	209.05	162.32	**63.16**
SHANK3	22	**92.14**	96.03	87.73	128.76	40.77	**23.37**
SHROOM4	X	**84.92**	199.93	92.8	102.24	73.7	**59.54**
SIL1	5	**167.93**	176.09	153.93	219	117.45	**52.09**
SLC16A2	X	**95.29**	191.86	89.71	102.77	75.19	**56.68**
SLC35A1	6	**94.52**	86.78	105.94	118.68	130.38	**42.01**
SLC35C1	11	**181.41**	178.83	157.21	220.06	118.75	**46.87**
SLC6A8	X	**6.11**	11.01	6.31	14.48	6.55	**2.61**
SNRPN	15	**69.64**	85.84	93.84	113.43	89.68	**31.19**
SOS1	2	**113.1**	114.37	125.16	130.21	145.75	**48.56**
SOX3	X	**59.69**	118.15	46.04	64.87	29.19	**31.19**
SUZ12	17	**71.62**	67.57	74.81	86.36	123.04	**34.85**
TCF4	18	**107.87**	120.1	130.09	145.84	113.73	**43.27**
TSC1	9	**191.94**	211.41	191.59	218.62	185.63	**65.81**
TSC2	16	**139.23**	161.14	155.1	205.55	78.3	**40.38**
TSPAN7	X	**76.43**	153.85	81.13	94.13	82.21	**52.58**
UBE2A	X	**58.21**	96.51	60.21	68.49	87.91	**44.96**
UBE3A	15	**57.82**	51	53.43	72.9	89.32	**23.11**
UPF3B	X	**38.43**	67.6	39.57	48.28	52.12	**31.34**
WHSC1	4	**149.14**	168.67	150.08	181.19	130.69	**51.42**

WHSC2	4	**107.28**	117.25	102.2	142.55	57.45	**28.88**
ZFHX1B	2	**105.2**	112.69	110.64	124.93	125.11	**39.93**
ZNF41	X	**80.14**	161.3	80.93	96.93	91.05	**55.53**
ZNF674	X	**3.54**	11.65	9.41	11.36	8.58	**3.15**
ZNF81	X	**80.94**	115.7	62.7	89.15	98.13	**59.43**

Supplementary Table 2 Primer sequences used for quantitative PCR (qPCR) to determine successful target enrichment. Genes in bold have capture probes on the array (positive controls), the others are negative controls.

Gene name	Target	Forward primer	Reverse primer
MECP2	Exon 1	CACCAGTTCCTGCTTTGATGT	CCCTAACATCCCAGCTACCAT
CREBBP	Exon 4	CACAAGTCCATTTGGACAGC	GTTGACCATGCTCTGTTTGC
CREBBP	Exon 5	CAGTGGGAATTGTACCCACAC	GAGCATGAAGCAGTAGAACCAG
NSD1	Exon 7	GTGAAGAGGAAAGCCTTCTAGC	AGAACTGGAGGCTCTTCTTTGG
MBL2	Exon 1	CCTGTTTCCATCACTCCCTCT	CACTGCAGGGCAGGTCTTTT
MBL2	Exon 4	AAGTGAAGGCCTTGTGTGTCA	AAGGCTTCCTCCTTGATGAGAT
BRCA1	Exon 10	CCCTTTGAGAGTGGAAGTGACA	CTGGGCTCCATTTAGACCTGA
DMD	Exon 20	TGCCAGTTGCTAAGTGAGAGAC	GCAGTAGTTGTCATCTGCTCCA
DMD	Exon 51	GGAAACTGCCATCTCCAAAC	CCAGTCGGTAAGTTCTGTCCA

Supplementary Table 3 Primer sequences (with M13 tail) used for amplification of targets for Sanger sequencing.

Gene name	Target	Forward primer	Reverse primer
DHCR7	exon4	TGTAAAACGAcggccagctcccacacagagcctcttagg	CAGGAAACAGCTATgacccccagacaaatggaaggactac
MED12	exon 6	TGTAAAACGACGGCcagcccagttgtggttctcttcatc	CAGGAAACAGCTATGACCaggggaccctgctctaacattt
ALG6	exon 6	TGTAAAACgacggccagctgggttcattgttgtaggtactg	CAGGAAACAGCTATGACCcttttccccaaacaaacacc
SCN8A	exon 17	TGTAAAACGACggccaggtctgtcacgtgaagtccattg	CAGGAAACAGCTATGACCctgtttctaggctgggaccttac

B4GALT1	exon 2	TGTAAAACGACggccagagtcctgtggcctgctaacttct	CAGGAAACAGCTATGACCgtctgtgaaatcactccccttc
B3GALTL	exon 5	TGTAAAACGACGGCCAGagtagtcaattcatacttatc Ttttcgg	CAGGAAACAGCTATGACCtgaggaaaaccacaccctct
B3GALTL	exon 6	TGTAAAACGACGGCCAGTtgccattctgtgtacccttc	CAGGAAACAGCTATGACCtggctcattataagctctgtcc
B3GALTL	exon 9	TGTAAAACGACGGCCAGTtgtgttctgctttcccttga	CAGGAAACAGCTATGACCagaatcagcagaagcccaaa
B3GALTL	exon 10	TGTAAAACGACGGCCagtatgccggaaatatgtttggt	CAGGAAACAGCTATGACCttgaaatggtgcaatgagga
B3GALTL	exon 13	TGTAAAAcgacgggcagtcagagtgggatgtaagaacca	CAGGAAACAGCTATGACCtcccagtgccagagacctac
B3GALTL	exon 15	TGTAAAACGACGGCCAGTggtagtgaagtaaagcag tccactt	CAGGAAACAGCTATGACCaagtcaggaagcacccaatg
NSD1	exon 23	TGTAAAACGACGGCCAGAACCTCCTGCTGACA CCAAC	CAGGAAACAGCTATGACCTGGGAACTGAGGTTTT CTCC
NSD1	exon 17	TGTAAAACGACGGCCAGTagcattggtcgatttttgtg	CAGGAAACAGCTATGACCgcccggctatttctgatctt
NSD1	exon 5	TGTAAACGACGGCCAGTAACCTCGTAAGCGCA TGAAC	CAGGAAACAGCTATGACCgggaaaagggcttctggtaa
TSC1	exon 23	TGTAAAACGACGGCCAGCCTAACCCCCTCTCA TTTACCT	CAGGAAACAGCTATGACCGGGACAAAACCAGACT TACCTG
RAB3GAP2	exon 35	TGTAAAACGACGGCCAGTTACAGAGTAGCAGC ACTGGAAAG	CAGGAAACAGCTATGACCCCAAGTTTCTTTGACT AGCCTCCT
EP300	exon 31	TGTAAAACGACGGCCAGGACTCAGCACCGATA ACTCAGACT	CAGGAAACAGCTATGACCCGGCTACTGCACAGTT CTTATG
CENPJ	exon 16	TGTAAAACGACggccaggtacaaacttgctccaccctct	CAGGAAACAGCTATGACCcaggtgtcacactgagtggttt

Supplementary Table 4 qPCR fold-enrichment values for four targets from different samples.

Sample /sex	Amplicon Name	PCR efficiency	Delta-CP	qPCR fold enrichment	% Average coverage per target	Coverage per nucleotide in all targets	Coverage for all exons
S-2 (female)	MECP2 amplicon 1	1.98	9.26	558	203	128	119
S-2 (female)	CREBBP amplicon 4	1.8	9.9	336	182		
S-2 (female)	CREBBP amplicon 5	1.79	10.39	428	182		
S-2 (female)	NSD1 amplicon 7	1.84	9.51	328	222		
S-8 (female)	MECP2 amplicon 1	1.98	8.37	304	56	44	41

S-8 (female)	CREBBP amplicon 4	1.8	8.56	153	49		
S-8 (female)	CREBBP amplicon 5	1.79	8.64	153	49		
S-8 (female)	NSD1 amplicon 7	1.84	7.95	127	73		
S-5 (male)	MECP2 amplicon 1	1.98	10.23	1000	99	120	111
S-5 (male)	CREBBP amplicon 4	1.8	11.5	862	184		
S-5 (male)	CREBBP amplicon 5	1.79	11.35	741	184		
S-5 (male)	NSD1 amplicon 7	1.84	10.59	637	208		
S-6 (male)	MECP2 amplicon 1	1.98	9.82	819	91	100	92
S-6 (male)	CREBBP amplicon 4	1.8	11.2	723	154		
S-6 (male)	CREBBP amplicon 5	1.79	10.5	452	154		
S-6 (male)	NSD1 amplicon 7	1.84	9.87	411	184		

Supplementary Table 5 CP values for non-targeted regions (negative controls) before and after array enrichment.

Sample ID	CP value (negative targets) before capture	CP value (negative targets) after capture	Delta-CP value
	MBL2 amplicon 1	**MBL2 amplicon 1**	
S-2	26	39	-13
S-8	26	38	-12
S-5	28	31	-3
S-6	26	31	-5
	MBL2 amplicon 4	**MBL2 amplicon 4**	
S-2	25	36	-11
S-8	26	27	-1
S-5	26	29	-3
S-6	25	37	-12
	BRCA1 amplicon 10	**BRCA1 amplicon 10**	
S-2	26	36	-10

S-8	26	40	-14
S-5	26	40	-14
S-6	27	40	-13
	DMD amplicon 10	**DMD amplicon 10**	
S-2	25	40	-15
S-8	26	39	-13
S-5	27	32	-5
S-6	26	31	-5
	DMD amplicon 20	**DMD amplicon 20**	
S-2	25	40	-15
S-8	26	28	-2
S-5	27	32	-5
S-6	28	40	-12

Supplementary Table 6 Single nucleotide polymorphism (SNPs) detected by Illumina sequencing and confirmed by SNP array data from two patients.

Sample id	Chromosome position	Gene name	SNP ID	Location	Ref. Sequence	Genotype array	Genotype Illumina	Sequence depth	Wild type	Mutant
S-7	chr22_48683439	ALG12	rs1321	exon	A	AG	AG	93	45	48
S-7	chrX_147856958	AFF2	rs16994895	Intron	T	GG	GG	95	0	95
S-7	chrX_147887801	AFF2	rs6641482	exon	A	GG	GG	66	0	66
S-7	chr4_123883877	BBS12	rs13135766	exon	G	GC	GC	156	84	72
S-7	chr13_30749059	B3GALTL	rs1409373	Intron	A	GG	GG	77	0	77
S-7	chr13_30803641	B3GALTL	rs912603	exon	G	GA	GA	153	73	80
S-7	chr3_3167927	CRBN	rs1620675	Intron	T	GG	GG	66	0	66
S-7	chr9_122348097	CDK5RAP2	rs10739564	Intron	T	CC	CC	56	0	56

S-7	chr9_122370634	CDK5RAP2	rs4837782	Intron	T	CC	CC	145	0	145
S-7	chr9_122209539	CDK5RAP2	rs2282168	Intron	C	GG	GG	59	0	59
S-7	chr22_39867180	EP300	rs6002267	Intron	G	TT	TT	206	0	206
S-7	chr22_39844101	EP300	rs4822005	Intron	G	AA	AA	16	0	16
S-7	chrX_53244068	JARID1C	rs2182285	Intron	A	GG	GG	11	0	11
S-7	chr12_25253819	KRAS	rs712	exon	A	AC	AC	27	15	12
S-7	chrX_28717690	IL1RAPL1	rs12690144	Intron	T	CC	CC	20	0	20
S-7	chrX_28717669	IL1RAPL1	rs6526807	Intron	A	GG	GG	35	0	35
S-7	chr8_6290433	MCPH1	rs2584	exon	G	GA	GA	166	72	94
S-7	chrX_70269142	MED12	rs10521349	Intron	T	CC	CC	49	0	49
S-7	chr13_44422083	NUFIP1	rs1175384	Intron	A	CC	CC	151	77	74
S-7	chr9_133371932	POMT1	rs10448341	Intron	G	AA	AA	19	0	19
S-7	chr3_12601516	RAF1	rs3729931	Intron	G	GA	GA	112	50	62
S-7	chr5_138484893	SIL1	rs11750382	Intron	G	GA	GA	100	49	51
S-7	chr5_138385110	SIL1	rs3749665	Intron	A	AG	AG	37	23	14
S-7	chr5_138414758	SIL1	rs3828600	Intron	C	CA	CA	153	74	79
S-7	chr2_199953490	SATB2	rs1348813	Intron	C	CG	CG	154	72	82
S-7	chr12_50449589	SCN8A	rs303809	Intron	G	CC	CC	108	0	108
S-7	chr12_50449515	SCN8A	rs303810	Intron	A	GG	GG	174	0	174
S-7	chr12_50469752	SCN8A	rs303816	Intron	C	TT	TT	20	0	20
S-5	chr1_195337065	ASPM	rs3762271	Exon	A	AC	AC	143	60	83
S-5	chr2_144878372	ZEB2	rs13009259	Intron	G	AA	AA	25	12	13
S-5	chr2_199845853	SATB2	rs2881208	Intron	T	CC	CC	30	0	30
S-5	chr4_57492171	REST	rs3796529	Intron	G	AG	AG	148	80	68
S-5	chr7_42054747	GLI3	rs846266	Exon	A	GG	GG	222	0	221
S-5	chr9_122211576	CDK5RAP2	rs2297454	Intron	A	AG	AG	42	24	18
S-5	chr9_133375257	POMT1	rs2296949	Exon	A	GG	GG	311	0	311
S-5	chr9_122211576	CDK5RAP2	rs2297454	Intron	T	TC	TC	42	18	24

S-5	chr9_134760121	TSC1	rs2809243	Exon	G	AA	AA	153	0	153
S-5	chr11_45789511	SLC35C1	rs1139266	Exon	G	AA	AA	133	0	133
S-5	chr11_66010661	DPP3	rs11550299	Exon	C	AC	AC	114	53	61
S-5	chr11_66028718	DPP3	rs1671063	Exon	A	GG	GG	109	0	109
S-5	chr11_66028813	DPP3	rs2305535	Exon	G	AG	AG	227	113	114
S-5	chr11_66038671	BBS1	rs2298806	Exon	G	AG	AG	210	106	104
S-5	chr11_66053939	BBS1	rs3816492	Exon	C	CT	CT	194	98	96
S-5	chr11_111229343	ALG9	rs10502151	Exon	G	AG	AG	115	57	58
S-5	chr12_25251108	KRAS	rs13096	Exon	A	AG	AG	53	27	26
S-5	chr12_50450056	SCN8A	rs303808	Intron	G	AG	AG	198	108	94
S-5	chr12_50470538	SCN8A	rs303815	Exon	A	AG	AG	115	47	68
S-5	chr13_30789746	B3GALTL	rs1041073	Exon	G	AA	AA	135	0	135
S-5	chr16_8849319	PMM2	rs2075827	Exon	A	CC	CC	210	0	210
S-5	chr17_7431901	MPDU1	rs4227	Exon	C	AA	AA	229	0	229
S-5	chr17_17637480	RAI1	rs11649804	Exon	C	CA	CA	237	130	107
S-5	chr18_51282486	TCF4	rs3760600	Intron	C	AC	AC	50	28	22
S-5	chr19_13890269	CC2D1A	rs2305776	Intron	A	AC	AC	21	11	10
S-5	chr20_10566574	JAG1	rs8708	Exon	A	GG	GG	201	0	201
S-5	chr20_10568275	JAG1	rs1051421	Exon	C	CT	CT	164	83	81
S-5	chr20_10581313	JAG1	rs6040055	Intron	A	AG	AG	206	98	108
S-5	chr20_30860197	DNMT3B	rs2424932	Exon	A	GG	GG	94	0	94
S-5	chr20_48986311	DPM1	rs2294902	Intron	A	GG	GG	63	0	63
S-5	chr22_49480384	SHANK3	rs13055562	Intron	G	AG	AG	106	45	61
S-5	chrX_5820574	NLGN4X	rs3810686	Exon	G	AA	AA	62	0	62
S-5	chrX_47212055	ZNF41	rs5905607	Exon	T	GG	GG	114	0	114
S-5	chrX_53980349	PHF8	rs7892782	Exon	T	CC	CC	45	0	45
S-5	chrX_54036020	PHF8	rs7061449	Intron	C	TT	TT	140	0	140
S-5	chrX_69590536	DLG3	rs2274309	Intron	T	CC	CC	36	0	36

S-5	chrX_69640819	DLG3	rs1044422	Exon	G	AA	AA	125	0	125
S-5	chrX_118853103	UPF3B	rs2239963	Intron	A	CC	CC	74	0	74
S-5	chrX_152945374	MECP2	rs2734647	Exon	T	CC	CC	105	0	105

Supplementary Table 7 Different variants detected by Illumina sequencing and confirmed by Sanger Sequencing

sample ID	chr. position	gene name	location	Ref. sequence	observed genotype	Change	sequence depth	wild type	variant	Sanger sequencing
S-2	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	17	0	17	AA
S-2	30789746	B3GALTL	Exon	G	AA	c.1108G>A	21	0	21	AA
S-2	30801834	B3GALTL	UTR	G	TT	c.*29G>T	94	0	94	TT
S-2	30719240	B3GALTL	Intron	C	CT	c.347+4C>T	19	5	14	CT
S-2	30801841	B3GALTL	UTR	A	GA	*36A>G	38	50	88	GA
S-2	30789743	B3GALTL	Exon	G	GA	c.1105G>A	11	3	8	GA
S-2	176571910	NSD1	Intron	C	CG	c.3796+108C>G	41	18	23	CG
S-2	176570797	NSD1	Exon	C	CG	c.2791C>G	66	28	38	CG
S-2	176653878	NSD1	Exon	G	GC	c.6903G>C	71	33	38	GC
S-2	152949971	MECP2	Exon	C	CT	c.538C>T	34	12	22	CT
S-3	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	23	0	23	AA
S-3	30719256	B3GALTL	Intron	C	GG	c.347+20C>G	15	0	15	GG
S-3	30789746	B3GALTL	Exon	G	AA	c.1108G>A	32	0	32	AA
S-3	30801834	B3GALTL	UTR	G	TT	c.*29G>T	93	0	93	TT
S-5	30719256	B3GALTL	Intron	C	CG	c.347+20C>G	113	33	80	CG
S-5	30719992	B3GALTL	Exon	T	CC	c.348T>C	13	0	13	CC
S-5	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	93	0	93	AA
S-5	30789746	B3GALTL	Exon	G	AA	c.1108G>A	135	0	135	AA
S-5	30801834	B3GALTL	UTR	G	TT	c.*29G>T	261	0	261	TT
s-5	30801841	B3GALTL	UTR	A	AG	c.*36A>G	156	33	123	GA
s-5	30719240	B3GALTL	Intron	C	CT	c.347+4C>T	107	31	76	CT
S-5	176571910	NSD1	Intron	C	GG	c.3796+108C>G	99	0	99	GG
S-6	30741415	B3GALTL	Intron	G	AA	c.660+1G>A	27	0	27	AA
S-7	30719256	B3GALTL	Intron	C	CG	c.347+20C>G	120	56	64	CG
S-7	30801841	B3GALTL	UTR	A	GA	c.*36A>G	160	83	77	GA
S-7	30719992	B3GALTL	Exon	T	CC	c.348T>C	17	0	17	CC
S-7	30746823	B3GALTL	Intron	A	AG	c.780+58A>G	218	94	124	AG
S-7	30748989	B3GALTL	Intron	G	GA	c.850+81G>A	168	84	84	AG
S-7	30789746	B3GALTL	Exon	G	GA	c.1108G>A	66	23	43	GA

S-7	30801834	B3GALTL	UTR	G	TG	c.*29G>T	131	74	57	TG
S-7	33125238	B4GALT1	Exon	C	CT	c.597C>T	67	31	36	CT
S-7	63644620	ALG6	Exon	T	CT	c.391T>C	146	85	61	CT
S-7	70832913	DHCR7	Intron	G	AG	c.99-4G>A	92	51	41	AG
S-7	70257894	MED12	Intron	A	CC	c.736-8A>C	45	0	45	CC
S-7	50449090	SCN8A	Exon	C	CT	c.3076C>T	125	57	68	CT
S-7	134761154	TSC1	UTR	T	del T	c.*289delT	52		52	c.*289delT
S-7	218389523	RAB3GAP2	UTR		insAAC	c.*866+827insAAC	44	0		c.*866+827insAAC
S-7	39904953	EP300	UTR	CAA	del CAA	*47_*49del	38	0	38	*47_*49del
S-7	24356236	CENPJ	Intron	CAA	del CAA	c.3704_15delCAA	44	0	44	c.3704_15delCAA
S-7	176653878	NSD1	Exon	G	GC	c.6903G>C	129	57	72	GC
S-7	176571910	NSD1	Intron	C	GG	c.3796+108C>G	109	0	109	GG
S-8	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	9	0	9	AA
S-8	30801834	B3GALTL	UTR	G	TT	c.*29G>T	43	0	43	TT
S-8	30719240	B3GALTL	Intron	C	TT	c.347+4C>T	30	0	30	TT

Supplementary Table 8 Number of variants detected in six different samples in UTR and introns.

Patient ID	Variants in introns	Variants in untranslated region
S-2	188	77
S-3	171	64
S-5	279	107
S-7	245	76
S-6	363	103
S-8	136	81

Supplementary Table 9 All variants detected in exons in six different samples (S-2, S-3, S-5, S-6, S-7, S-8).

Chromosome position	Gene name	Variant type	Variant	SNP ID
chrX_147842900	AFF2	Silent	c.1488G>A	rs12011040
chr22_48687480	ALG12	Silent	c.885A>G	rs8135963
chr1_195337438	ASPM	Silent	c.7566A>G	rs1412640
chr1_195358160	ASPM	Silent	c.3579T>A	rs4915337
chr1_195360653	ASPM	Silent	c.3138G>A	rs6676084
chr1_195379156	ASPM	Silent	c.849C>T	rs6677082
chr1_195337330	ASPM	Silent	c.7674C>T	rs41308365
chr1_195339043	ASPM	Silent	c.5961A>G	rs41310925
chr1_195340555	ASPM	Silent	c.4449A>G	rs2878749

chr1_195375569	ASPM	Silent	c.1977T>C	rs17550662
chr1_195337399	ASPM	Silent	c.7605G>A	rs10922162
chr13_30801679	B3GALTL	Silent	c.1371A>G	
chr13_30719992	B3GALTL	Silent	c.348T>C	rs4943266
chr9_33125238	B4GALT1	Silent	c.597C>T	rs1065765
chr4_123883877	BBS12	Silent	c.1380G>C	rs13135766
chr4_123883907	BBS12	Silent	c.1410C>T	rs13135445
chr4_123884369	BBS12	Silent	c.1872A>G	rs13102440
chr4_123883559	BBS12	Silent	c.1062G>C	rs34296401
chr4_123883697	BBS12	Silent	c.1200G>A	rs309371
chr4_123883706	BBS12	Silent	c.1209G>A	rs17006092
chr4_123883895	BBS12	Silent	c.1398C>T	rs2292493
chr8_86576655	CA2	Silent	c.562T>C	rs703
chr19_13891689	CC2D1A	Silent	c.1281T>C	rs10410239
chr9_122202874	CDK5RAP2	Silent	c.5418C>T	rs3739822
chr9_122222023	CDK5RAP2	Silent	c.4041G>A	rs6478475
chr9_122260650	CDK5RAP2	Silent	c.2274T>C	rs2501727
chr9_122202874	CDK5RAP2	Silent	c.5418C>T	rs3739822
chr13_24364955	CENPJ	Silent	c.3042A>G	rs3742165
chr16_3717837	CREBBP	Silent	c.7212A>G	rs55916120
chr11_70824339	DHCR7	Silent	c.1158T>C	rs760241
chr11_70830109	DHCR7	Silent	c.438T>C	rs949177
chr11_70832801	DHCR7	Silent	c.207T>C	rs1790334
chr11_70832819	DHCR7	Silent	c.189G>A	rs1044482
chr11_70832777	DHCR7	Silent	c.231C>T	rs4316537
chr11_70824225	DHCR7	Silent	c.1272C>T	rs909217
chr20_30850008	DNMT3B	Silent	c.1572T>C	rs6058891
chr20_30850110	DNMT3B	Silent	c.1674T>C	rs2424922
chr11_118484236	DPAGT1	silent	c.16A>G	rs6589717
chr22_39880985	EP300	silent	c.3183T>A	rs20552
chr22_39903214	EP300	silent	c.5553T>C	
chr5_60236422	ERCC8	silent	c.435T>C	rs4647100
chr4_1777692	FGFR3	silent	c.1959G>A	rs7688609
chr4_1773502	FGFR3	silent	c.882T>C	rs2234909
chr7_41971125	GLI3	silent	c.4071C>T	rs34089404
chr7_41972361	GLI3	silent	c.2835G>C	rs61758978
chr7_42046290	GLI3	silent	c.900C>T	rs35961850
chr7_42054757	GLI3	silent	c.537C>T	rs3898405
chrX_122364958	GRIA3	silent	c.1200T>C	rs502434
chr6_102610010	GRIK2	silent	c.2424G>A	rs2227283
chr11_524242	HRAS	silent	c.81T>C	rs12628
chr20_10568275	JAG1	silent	c.3528C>T	rs1051421
chr20_10568386	JAG1	silent	c.3417T>C	rs1051419

chr20_10581237	JAG1	silent	c.765C>T	rs1131695
chr20_10573804	JAG1	silent	c.2214A>C	rs1801140
chr20_10585057	JAG1	silent	c.744A>G	rs10485741
chr20_10601469	JAG1	silent	c.267G>A	rs1051415
chr20_10587222	JAG1	silent	c.588C>T	rs1801138
chr12_25259729	KRAS	silent	c.483G>A	rs4362222
chr12_25254044	KRAS	silent	c.519T>C	rs1137282
chrX_43475980	MAOA	silent	c.891G>T	rs6323
chrX_43488335	MAOA	silent	c.1410T>C	rs1137070
chr8_6290433	MCPH1	silent	c.1782G>A	rs2584
chr8_6466586	MCPH1	silent	c.2418C>A	rs2912016
chr8_6259807	MCPH1	silent	c.228G>T	rs2305022
chr8_6466394	MCPH1	silent	c.2226C>T	rs2912010
chrX_70266672	MED12	silent	c.3930A>C	rs5030619
chrX_70277813	MED12	silent	c.6276G>A	
chr15_72976983	MPI	silent	c.1131A>G	rs1130741
chr17_26577611	NF1	silent	c.2034G>A	rs2285892
chr17_26532901	NF1	silent	c.702G>A	rs1801052
chr17_26507234	NF1	silent	c.168C>T	rs17881168
chr5_176569488	NSD1	silent	c.675C>T	rs1363405
chr5_176569755	NSD1	silent	c.1749G>A	rs3733874
chr5_176653804	NSD1	silent	c.6829T>C	rs28580074
chr5_176653878	NSD1	silent	c.6903G>C	rs11740250
chr9_133377309	POMT1	silent	c.1113C>T	rs3739494
chr9_133385395	POMT1	silent	c.1758G>A	rs34954751
chrX_48644671	PQBP1	silent	c.510G>A	
chr4_119422614	PRSS12	silent	c.2553A>C	
chr4_119456796	PRSS12	silent	c.1281G>A	rs2292597
chr1_218397295	RAB3GAP2	silent	c.3495G>A	rs11547779
chr17_17638979	RAI1	silent	c.1992G>A	rs8067439
chr17_17637824	RAI1	silent	c.837G>A	rs11078398
chr4_57471795	REST	silent	c.234G>T	rs61748752
chr4_57492946	REST	silent	c.3165G>A	rs2227901
chr17_26322577	RNF135	silent	c.360G>T	rs7224960
chrX_20114382	RPS6KA3	silent	c.798C>A	rs12009120
chr12_50367232	SCN8A	silent	c.576C>T	rs4761829
chr12_50470538	SCN8A	silent	c.4509T>C	rs303815
chr12_50487009	SCN8A	silent	c.5472C>A	rs60637
chrX_50367414	SHROOM4	silent	c.3468A>G	rs3747282
chr5_138484714	SIL1	silent	c.153A>G	rs3088052
chr18_51046529	TCF4	silent	c.1941A>G	rs8766
chr9_134762538	TSC1	silent	c.2829C>T	rs4962081
chr9_134772042	TSC1	silent	c.1335A>G	rs7862221

chr16_2078585	TSC2	silent	c.5397G>C	rs1051771
chr16_2076341	TSC2	silent	c.4809C>T	
chr16_2078270	TSC2	silent	c.5202T>C	rs1748
chr16_2074493	TSC2	silent	c.4269G>A	rs45438898
chrX_147856140	AFF2	missense	c.3040G>A	
chr22_48683892	ALG12	missense	c.1177A>G	rs3922872
chr1_63654140	ALG6	missense	c.911C>T	rs4630153
chr1_63644620	ALG6	missense	c.391T>C	rs35383149
chr11_77501439	ALG8	missense	c.803G>A	rs61995925
chr11_111229343	ALG9	missense	c.352G>A	rs10502151
chr1_195337524	ASPM	missense	c.7480T>C	rs964201
chr1_195327709	ASPM	missense	c.9395T>G	rs36004306
chr1_195337065	ASPM	missense	c.7939C>A	rs3762271
chr1_195337320	ASPM	missense	c.7684A>G	rs41310927
chr1_195339155	ASPM	missense	c.5849C>T	
chr13_30789743	B3GALTL	missense	c.1105G>A	rs34638481
chr13_30789746	B3GALTL	missense	c.1108G>A	rs1041073
chr11_66038671	BBS1	missense	c.378G>A	rs2298806
chr12_75264280	BBS10	missense	c.1616C>T	rs35676114
chr4_123883654	BBS12	missense	c.1157G>A	rs309370
chr4_123883896	BBS12	missense	c.1399G>A	rs13135778
chr16_55106002	BBS2	missense	c.209G>A	rs4784677
chr16_55102676	BBS2	missense	c.367A>G	rs11373
chrX_79830225	BRWD3	missense	c.3863A>G	rs3122407
chr19_13899791	CC2D1A	missense	c.2402C>T	rs2305777
chr19_13891753	CC2D1A	missense	c.1345G>A	
chr9_122210554	CDK5RAP2	missense	c.4618G>C	rs4837768
chr9_122330857	CDK5RAP2	missense	c.865G>C	rs4836822
chr9_122245733	CDK5RAP2	missense	c.3134G>C	rs3780679
chr9_122245802	CDK5RAP2	missense	c.3065G>A	rs34523498
chr13_24377541	CENPJ	missense	c.2635T>G	rs17402892
chr13_24384911	CENPJ	missense	c.253C>A	rs9511510
chrX_69582011	DLG3	missense		c.235G>A
chr11_118472968	DPAGT1	missense	c.1177A>G	rs643788
chr11_66010661	DPP3	missense	c.435G>T	rs11550299
chr11_66028813	DPP3	missense	c.2033G>A	rs2305535
chr22_39877954	EP300	missense	c.2989A>G	rs20551
chr9_107406555	FKTN	missense	c.608G>A	rs34787999
chr9_107437316	FKTN	missense	c.1336A>G	rs41313301
chrX_106733095	FRMPD3	missense	c.5269C>G	
chrX_106733134	FRMPD3	missense	c.5308C>G	
chr7_42054747	GLI3	missense	c.547A>G	rs846266
chr7_41972203	GLI3	missense	c.2993C>T	rs929387

chr20_10570501	JAG1	missense	c.2612C>G	rs35761929
chr8_6283958	MCPH1	missense	c.513G>T	rs2442513
chr8_6289591	MCPH1	missense	c.940G>C	rs930557
chr8_6289826	MCPH1	missense	c.1175A>G	rs2515569
chr8_6466450	MCPH1	missense	c.2282C>T	rs1057090
chr8_6487952	MCPH1	missense	c.2482C>T	rs1057091
chr8_6325714	MCPH1	missense	c.2045C>A	rs12674488
chr17_7431534	MPDU1	missense	c.685G>A	rs10852891
chr2_15999836	MYCN	missense	c.199T>G	
chr17_26725232	NF1	missense	c.8453C>A	
chr5_176569846	NSD1	missense	c.1840G>T	rs3733875
chr5_176570182	NSD1	missense	c.2176T>C	rs28932178
chr5_176570797	NSD1	missense	c.2791C>G	
chr13_44461464	NUFIP1	missense	c.108C>G	rs1140993
chrX_67569473	OPHN1	missense	c.115G>A	rs41303733
chr9_133375257	POMT1	missense	c.752A>G	rs2296949
chr9_133376602	POMT1	missense	c.979G>A	rs4740164
chr4_119422669	PRSS12	missense	c.2498G>A	rs17594503
chr4_119422617	PRSS12	missense	c.2550T>G	
chr4_119493160	PRSS12	missense	c.164G>C	rs13119545
chr1_218391338	RAB3GAP2	missense	c.4060A>G	rs59190330
chr1_218397828	RAB3GAP2	missense	c.3275G>C	rs2289189
chr17_17637256	RAI1	missense	c.269G>C	rs3803763
chr17_17637480	RAI1	missense	c.493C>A	rs11649804
chr17_17639523	RAI1	missense	c.2536T>G	
chr17_17647830	RAI1	missense	c.5601T>C	rs3818717
chr17_17641713	RAI1	missense	c.4726C>T	
chr4_57492171	REST	missense	c.2390C>T	rs3796529
chr4_57491857	REST	missense	c.2076G>T	rs2227902
chr17_26322430	RNF135	missense	c.213C>G	rs7225888
chr17_26322539	RNF135	missense	c.322T>C	rs7211440
chr12_50401628	SCN8A	missense	c.1667T>G	
chr12_50401630	SCN8A	missense	c.1669T>C	
chr12_50449090	SCN8A	missense	c.3076C>T	
chrX_73558294	SLC16A2	missense	c.319T>C	rs6647476
chr15_22770605	SNRPN	missense	c.694T>C	rs705
chr9_134761574	TSC1	missense	c.3364G>A	
chr9_134776725	TSC1	missense	c.965T>C	rs1073123
chrX_152949971	MECP2	nonsense	c.538C>T	

Supplementary Table 10 All variants detected in B3GALTL gene in sample S-6. Variants at the first exons (1-7) can be heterozygous or homozygous while all variants at the rest of the gene (8-15) have only a homozygous genotype.

Chromosome position	location	Ref. sequence	observed genotype Illumina	Wt	Variant	Mutation
chr13_30694969	intron 2	G	CC	0	41	c.121-120G>C
chr13_30719240	intron 5	C	CT	59	42	c.347+4C>T
chr13_30719469	intron 5	A	GG	0	9	c.347+233A>G
chr13_30719992	exon 6	T>	CC	0	11	c.348T>C
chr13_30733375	intron 7	G	GA	33	32	c.596+156G>A
chr13_30741415	intron 8	G	AA	0	27	c.660+1G>A
chr13_30746535	intron 8	G	AA	0	37	c.661-111G>A
chr13_30741664	intron 8	A	GG	0	14	c.660+250A>G
chr13_30746823	intron 9	A	GG	0	91	c.780+58A>G
chr13_30748714	intron 9	G	CC	0	10	c.781-125G>C
chr13_30749059	intron 10	A	GG	0	43	c.850+151A>G
chr13_30758769	intron 11	G	CC	0	9	c.965-88G>C
chr13_30757039	intron 11	T	AA	0	23	c.964+141T>A
chr13_30759059	intron 12	G	AA	0	37	c.1064+103G>A
chr13_30789561	intron 12	T	CC	0	13	c.1065-142T>C