Experiences with array-based sequence capture; towards clinical applications

Johan T den Dunnen, Rowida Almomani, Jaap van Der Heijden, Yavuz Ariyurek, Yuchig Lai, Michiel van Galen, Martijn H. Breuning

Experiences with array-based sequence capture; towards clinical applications

Rowida Almomani[1], Jaap van der Heijden[1], Yavuz Ariyurek[1,2], Yuchig Lai[2], Michiel van Galen [1,2], Martijn H. Breuning[1] and Johan T. den Dunnen[1,2]

[1]Center for Human and Clinical Genetics and [2] Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands

Running title: Sequencing; target enrichment using array capture

Keywords: capture array, heterogeneous disorders, sequencing

Address for correspondence and reprints: Prof. Dr. Johan den Dunnen, Center for Human and Clinical Genetics and Leiden Genome Technology Center, Leiden University Medical Center, Postzone S4-P, P.O. Box 9600, 2300 RC, Leiden, The Netherlands. Telephone: +31-71-5269501, fax: +31-71-526-8285, E-mail: ddunnen@HumGen.nl

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 NimbleGen 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 which detects sequence variants, positions them in relation to the gene, checks for presence in databases (e.g. dbSNP) and predicts the potential consequences at the level of RNA splicing and protein translation. In the samples analyzed all known variants were reliably detected, incl. 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 (CNVs), incl. 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 re-sequencing of many disease-associated genes is required in order 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 re-sequencing 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 was 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 way.[13, 14, 15] We have used custom high-density microarrays (NimbleGen) for the enrichment of 112 distinct genes potentially involved in mental retardation and congenital malformation, followed by sequencing on the Illumina Genome Analyzer I platform.

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.

Material 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, carry a large deletion spanning exons 8 to 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 mental retardation 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, S-8.

Exon array design

Microarrays with 385K probe capacity (NimbleGen) were used to capture all exons, splice site and immediately adjacent intron sequence of 112 human genes. Based on searches in OMIM and literature we selected 112 human genes known to cause MR either as part of a known syndrome or in isolation Sup. Table 1. Primary sequence data from all exons was extracted from NCBI's genome (Build 36). Microarrays were designed by NimbleGen with long oligonucleotide probes (54-99 nucleotides) which span each target region, overlapped and shifted on 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 non–specific 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).[16] Genomic DNA (20-10μg) 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 hours. After hybridization and washing, the DNA fragments bound to

the array were eluted, using 300 µl of the elution buffer (Qiagen) on each array. A gasket (Agilent) was applied and placed on the thermal elution device (home-made) for 20 min at 95 ºC. We repeated this process once by adding 200 µl 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, S-8) before and after array enrichment. The primers amplified five loci from *MBL2*, *DMD*, and *BRCA1* (100bp) 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) Sup. Table 2. All primers for qPCR were designed using Primer 3 (http://frodo.wi.mit.edu/).

The qPCR assays were done in triplicate in the Lightcycler using 384-well plates (Roche) in 10 µl total volume: 5 µl of 2x SYBR Green master Rox (Roche), 0.25 µl of each primer (10 pmol/µl), 2 µl of DNA template and 2.5 µl of ultra–pure water. The thermo–cycling protocol was done as follows: 10 min at 95 ºC, 45 cycles of 10 s at 95 ºC, 30 s at 60 ºC, 20 s at 72 ºC, and 5 min at 72 ºC, followed by melting curve analysis in order to determine the specific and non–specific amplified products and other artifacts that might interfere with the 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 Δ CP (i.e. 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 = Δ CP (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 done 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), nucleotide level and per probe region. SNP calling was performed by searching for nucleotides discordant with the reference genome with 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. Then all variants were checked for their presence in known databases, e.g. dbSNP. Perl scripts were designed to predict the potential consequences at the level of RNA splicing and protein translation based on 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 upon request..

Sanger Sequencing

Twenty-one 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 the Sup. Table 3.

Results

The methodology used starts with fragmentation of the genomic DNA. Linker and primer addition can then be performed either prior to 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 prior to array capture. Initially experiments were performed using 20µg genomic DNA, later we reduced this to 10µg. We used qPCR, comparing targeted (four positive controls) and non-targeted regions (five negative controls), to check successful array enrichment as well as to estimate the fold-enrichment obtained, Sup. Tables 4 and 5 for examples. Since enrichment varies significantly from locus to locus, we tested multiple loci to get an accurate estimate. Samples where qPCR did not indicate clear enrichment (> 100x) 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. Since the fold-enrichments determined by qPCR correlate positively with the average sequence depth obtained, we conclude that qPCR provides an effective and cost- saving check for successful enrichment, examples are listed in Sup. Tables 4 and 5.

Sequence data

The custom arrays used contained 112 different human genes that are known to be or potentially involved in mental retardation (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 1 or 2 mismatches (M1, M2) (Fig 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 (>20x) to detect any variants that were present.

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) Sup. 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 Sup. Table 7. Similarly we detected a homozygous change in the *B3GALTL* gene in a Peter's Plus patient (c.660+1G>A, Sup. Table 7, see below).
Next, we 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 Sup. Table 7. The analysis of the variants found in all 112 genes of the patients did not reveal a clear cause of their mental retardation Sup.Table 8 and 9.

Copy number variations (CNV)

Changes that cannot be easily detected using the sequence itself include deletions and duplications (CNVs). However, such variants can be expected to give quantitative changes in coverage. To see whether overall coverage can be used to detect quantitative changes, we first analysed the 39 genes located on the X-chromosome. Indeed, when coverage was normalized using autosomal genes (Fig.2A), samples from female showed a clearly higher X-chromosome coverage compared to male samples (Fig.2B). Furthermore, as expected, the gene on the Y-chromosome (*NLGN4Y*) gave no coverage in the female sample (Fig.2B). To determine the sensitivity of our method for detecting smaller CNVs, we carefully analysed 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 since, as compared to other samples, we observed a significantly lower average coverage for the *B3GALTL* gene (53x versus 155x, 150x, 140x) (Fig.2C). In addition, while the splice site mutation in exon 8 was detected in the "homozygous" state (like all nine variants downstream), we observed variants in the first exons (1-7) also in heterozygous state Sup. Table 10. These data show that not only have we obtained an excellent specificity of the capture process but that we have also been able to distinguish between male and female samples.

Discussion

Array-based genomic selection offers several advantages for large-scale targeted DNA isolation over other approaches like 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. Since 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 mental retardation. 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 Sup. 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 5 samples analyzed could be confirmed using Sanger sequencing Sup. 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 Sup. Table 1, Fig.2A. Since 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 the hybridization, washing conditions, and potentially re-using the capture array. Furthermore coverage is influenced by array design, including probe sequence (melting temperature, GC content), probe density, and spacing Sup. Table 1. Our data show that AT-rich regions (>55%), regions with an overall low probe density (<3) and small exons (on average 90bp) yield a low coverage, which also varies significantly between experiments. For a second-generation capture-array the results obtained could be used to change probe density, i.e. decreased in well-covered and increased in low-covered regions.

Our data shows that longer reads (50 bp) improves accuracy and selectivity of read mapping to the reference genome which influenced the SNP calling by having less false positives and slightly better coverage.

Since CNVs (deletions / duplications) are a significant cause in the etiology of mental retardation[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 which gave no coverage in females (Fig.2B), while the 39 X-linked genes when compared to the 69 autosomal genes yielded overall 50% lower coverage in the male samples (Fig.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 (Fig. 2D). An algorithm for detecting local deviations from the average coverage is currently under development.

Regarding probe design (performed by NimbleGen) 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 while for diagnostic applications quality assurance requires that sequences are obtained in forward and reverse orientation. Sequencing this one strand in both directions is partly fooling one self. 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 re-used 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 e.g. 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), i.e. directly 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, since more flanking DNA is captured. Especially for array-based capture, due to 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 (i.e. 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 next generation sequencing 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 where 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. Based on initial experiences and publications, we expect that array-capture will quickly be 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.

ACKNOWLEDGEMENTS

We would like to 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 manuscript. 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: 5463-7.

2- Chelly J, Khelfaoui M, Francis F, Chérif B, Bienvenu T: Genetics and pathophysiology of mental retardation. Eur J Hum Genet 2006; 14: 701-13.

3- Szigeti K, Lupski JR: Charcot-Marie-Tooth disease. Eur J Hum Genet 2009; 17: 703-10.

4- Paul M, Zumhagen S, Stallmeyer B, Koopmann M, Spieker T, Schulze-Bahr E: Genes causing inherited forms of cardiomyopathies. A current compendium. 2009; 34: 98-109.

5- Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet 2006; 368: 1795-809.

6- Muhle R, Trentacoste SV, Rapin I. The genetics of autism. Pediatrics 2004; 113: 472-86.

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: 189-96.

8- Freeze H: Genetic defects in the human glycome. *Nat Rev Genet.* 2006; 7: 537–51.

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; 26:1135-45.

13- Albert TJ, Molla MN, Muzny DM *et al:* Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007; 4: 903-5.

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: 907-9.

15- Hodges E, Xuan Z, Balija V *et al:* Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39: 1522-7.

16- Roche NimbleGen. 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: 1913-25.

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: 9387-92.

22- Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M: Direct genomic selection. Nat Methods 2005; 2: 63-9.

23- Tewhey R, Warner JB, Nakano M *et al:* Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009; 27: 1025-31.

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: 241-8.

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: 182-9.

Table and Figure legends:

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

Figure 1. Detection of sequence variants. 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. "x n" behind the read indicates how many identical reads were obtained.

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 0 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 to 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) Average coverage per nucleotide for the second half (exons 8-15) of *B3GALTL* gene, the patient sample shows lower average coverage for this region compared with wild type. del = deletion, wt = wild type.

Table 1

Sample ID, sex	Total reads $x10^3$	Reads passing QC filter $x10^3$	Total number of reads mapped $x10^3$	MM0 reads $x10^3$	MM1 reads $x10^3$	MM2 reads $x10^3$	coverage per nucleotide	%of Nucleotides were covered >=8 times	% of nucleotides were covered 0 times	Read length	Array re-used
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%	2.70%	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

Figure 1

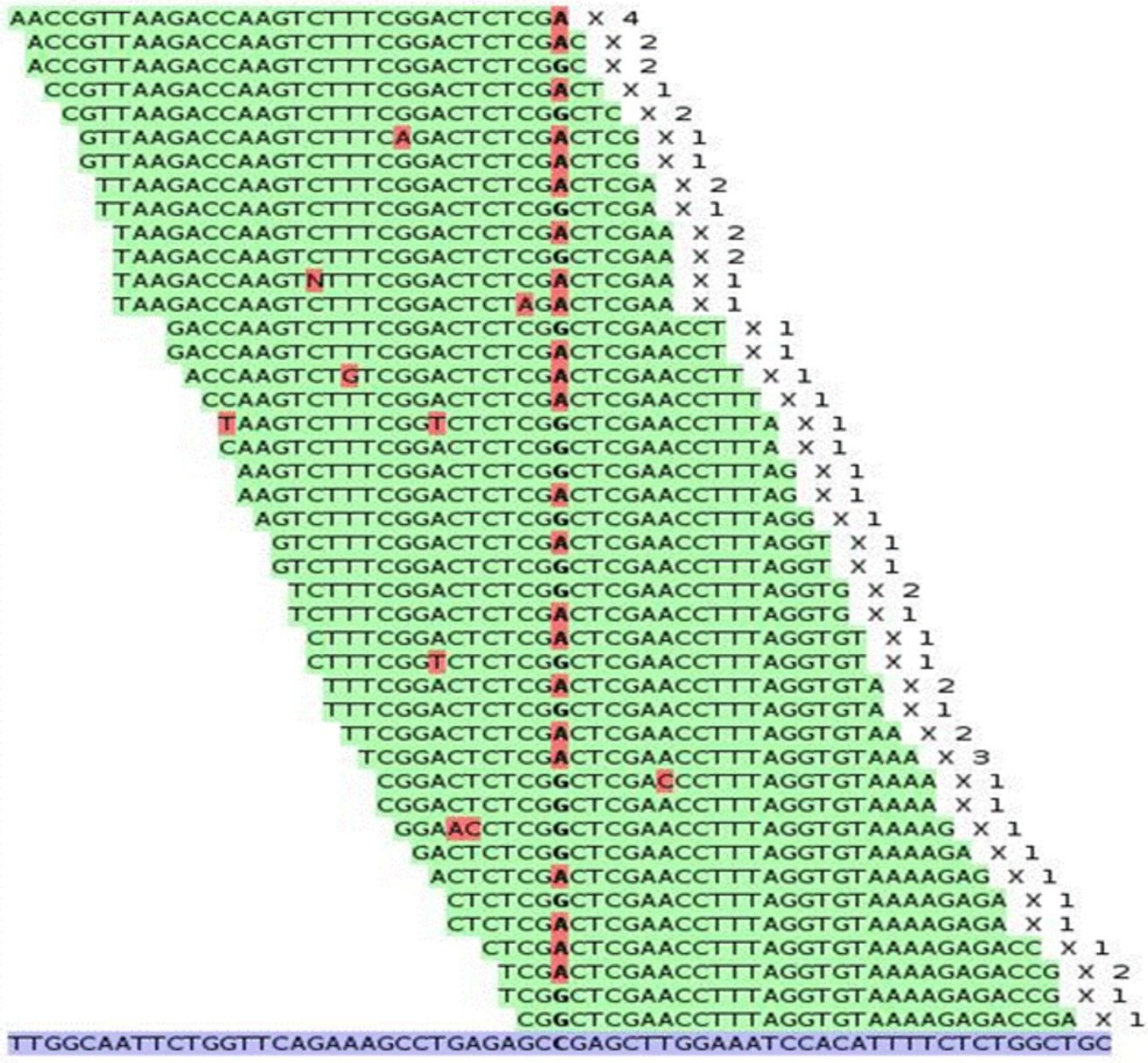

AACCGTTAAGACCAAGTCTTTCGGACTCTCGA X 4
ACCGTTAAGACCAAGTCTTTCGGACTCTCGAC X 2
ACCGTTAAGACCAAGTCTTTCGGACTCTCGGC X 2
CCGTTAAGACCAAGTCTTTCGGACTCTCGACT X 1
CGTTAAGACCAAGTCTTTCGGACTCTCGGCTC X 2
GTTAAGACCAAGTCTTTCAGACTCTCGACTCG X 1
GTTAAGACCAAGTCTTTCGGACTCTCGACTCG X 1
TTAAGACCAAGTCTTTCGGACTCTCGACTCGA X 2
TTAAGACCAAGTCTTTCGGACTCTCGGCTCGA X 1
TAAGACCAAGTCTTTCGGACTCTCGACTCGAA X 2
TAAGACCAAGTCTTTCGGACTCTCGGCTCGAA X 2
TAAGACCAAGTNTTTCGGACTCTCGACTCGAA X 1
TAAGACCAAGTCTTTCGGACTCTAGACTCGAA X 1
GACCAAGTCTTTCGGACTCTCGGCTCGAACCT X 1
GACCAAGTCTTTCGGACTCTCGACTCGAACCT X 1
ACCAAGTCTGTCGGACTCTCGACTCGAACCTT X 1
CCAAGTCTTTCGGACTCTCGACTCGAACCTTT X 1
TAAGTCTTTCGGTCTCTCGGCTCGAACCTTTA X 1
CAAGTCTTTCGGACTCTCGGCTCGAACCTTTA X 1
AAGTCTTTCGGACTCTCGGCTCGAACCTTTAG X 1
AAGTCTTTCGGACTCTCGACTCGAACCTTTAG X 1
AGTCTTTCGGACTCTCGGCTCGAACCTTTAGG X 1
GTCTTTCGGACTCTCGACTCGAACCTTTAGGT X 1
GTCTTTCGGACTCTCGGCTCGAACCTTTAGGT X 1
TCTTTCGGACTCTCGGCTCGAACCTTTAGGTG X 2
TCTTTCGGACTCTCGACTCGAACCTTTAGGTG X 1
CTTTCGGACTCTCGACTCGAACCTTTAGGTGT X 1
CTTTCGGTCTCTCGGCTCGAACCTTTAGGTGT X 1
TTTCGGACTCTCGACTCGAACCTTTAGGTGTA X 2
TTTCGGACTCTCGGCTCGAACCTTTAGGTGTA X 1
TTCGGACTCTCGACTCGAACCTTTAGGTGTAA X 2
TCGGACTCTCGACTCGAACCTTTAGGTGTAAA X 3
CGGACTCTCGGCTCGACCCTTTAGGTGTAAAA X 1
CGGACTCTCGGCTCGAACCTTTAGGTGTAAAA X 1
GGAACCTCGGCTCGAACCTTTAGGTGTAAAAG X 1
GACTCTCGGCTCGAACCTTTAGGTGTAAAAGA X 1
ACTCTCGACTCGAACCTTTAGGTGTAAAAGAG X 1
CTCTCGGCTCGAACCTTTAGGTGTAAAAGAGA X 1
CTCTCGACTCGAACCTTTAGGTGTAAAAGAGA X 1
CTCGACTCGAACCTTTAGGTGTAAAAGAGACC X 1
TCGACTCGAACCTTTAGGTGTAAAAGAGACCG X 2
TCGGCTCGAACCTTTAGGTGTAAAAGAGACCG X 1
CGGCTCGAACCTTTAGGTGTAAAAGAGACCGA X 1
TTGGCAATTCTGGTTCAGAAAGCCTGAGAGCCGAGCTTGGAAATCCACATTTTCTCTGGCTGC

Figure 2

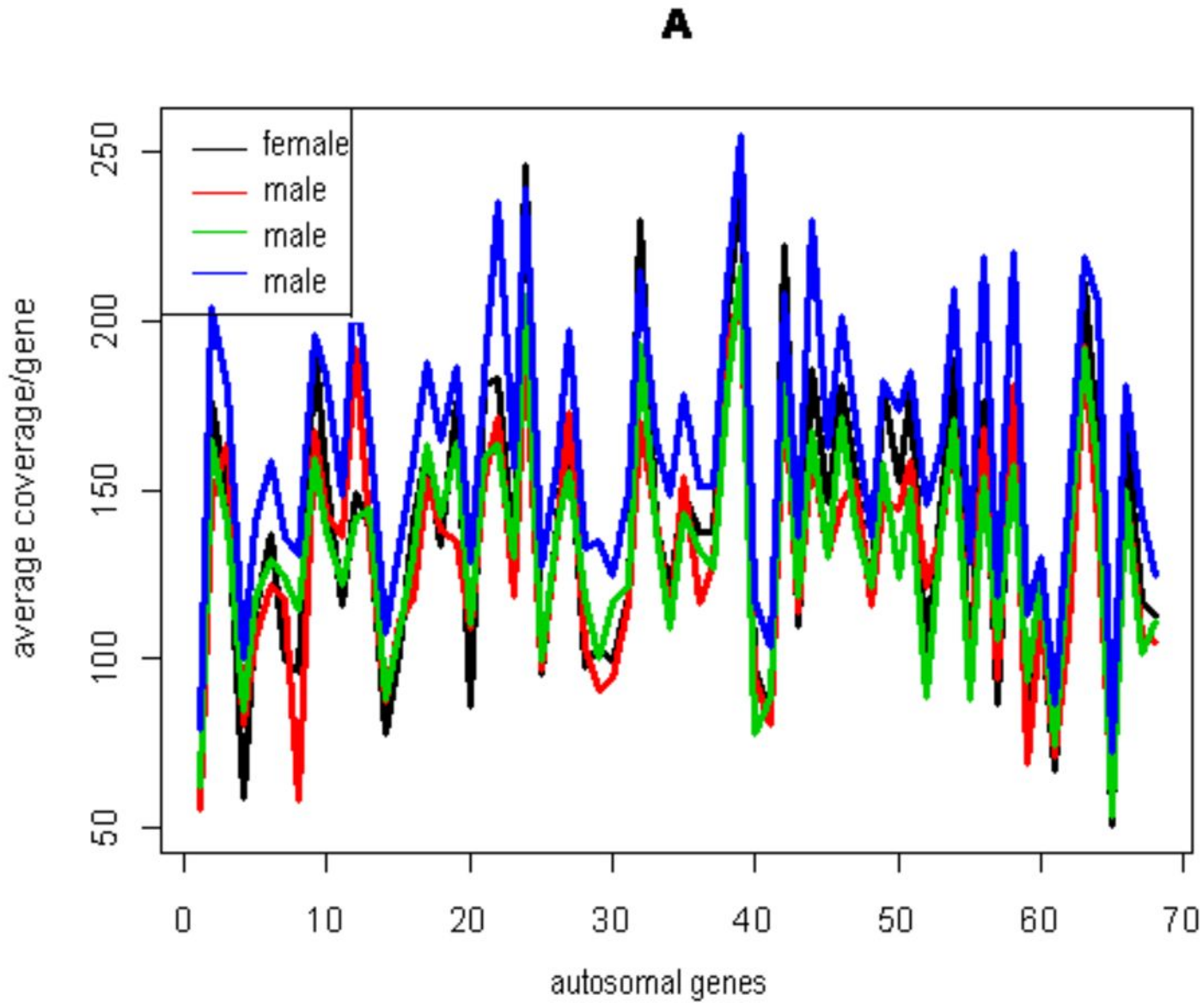

A
female
male
male
male
average coverage/gene
50
100
150
200
250
0
10
20
30
40
50
60
70
autosomal genes

Figure 2

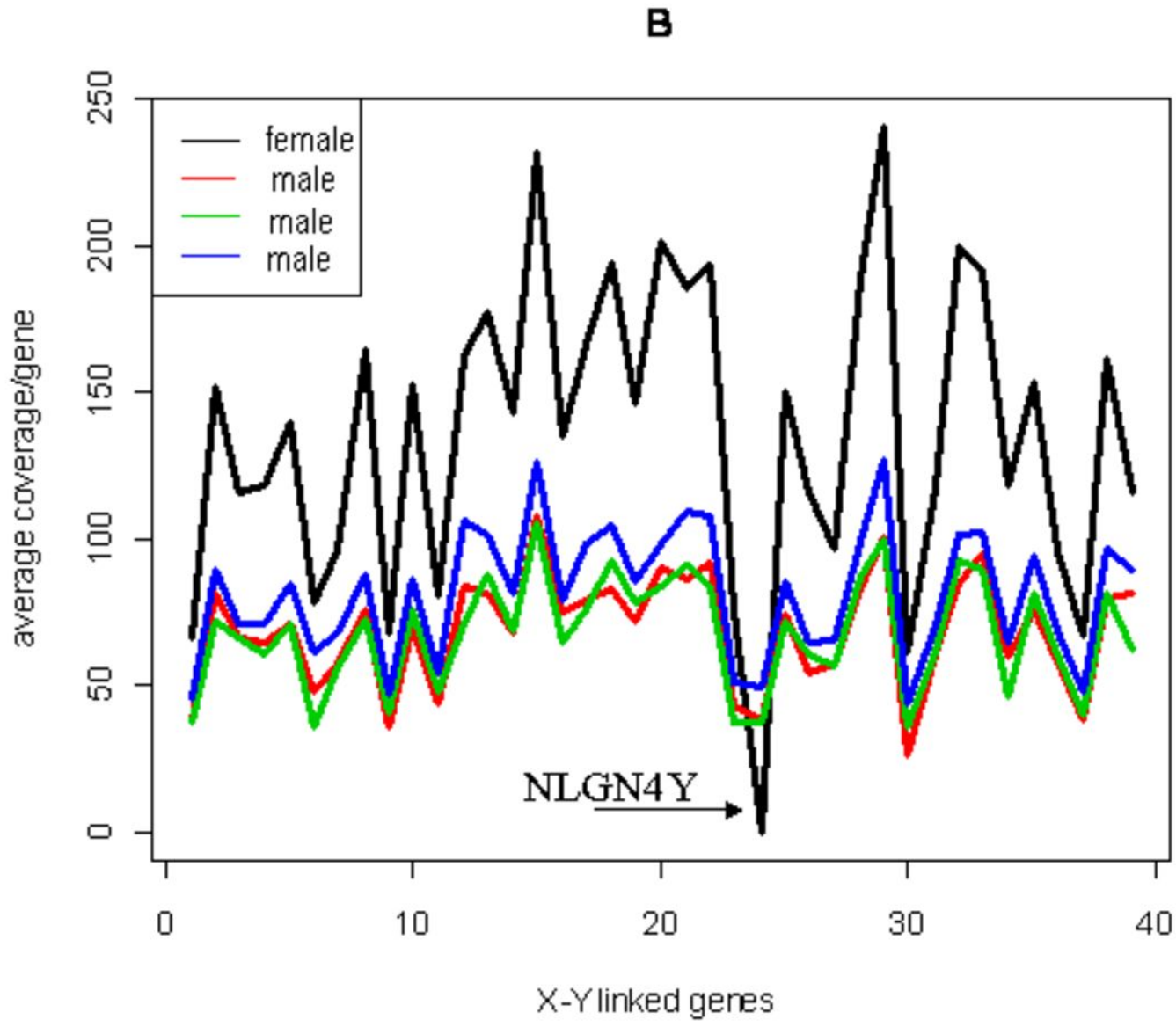

B
female
male
male
male
average coverage/gene
0
50
100
150
200
250
NLGN4Y
0
10
20
30
40
X-Y linked genes

Figure 2

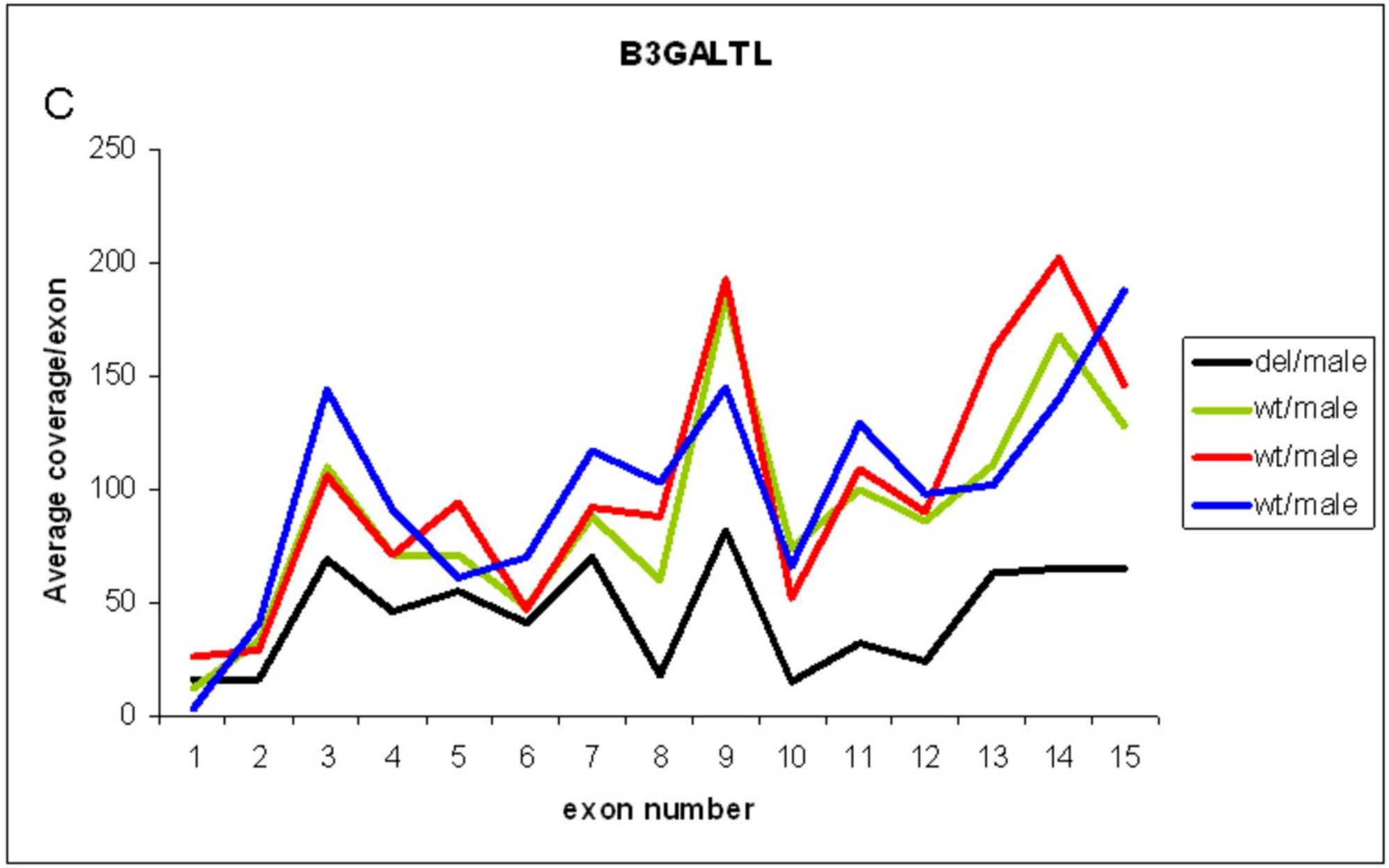
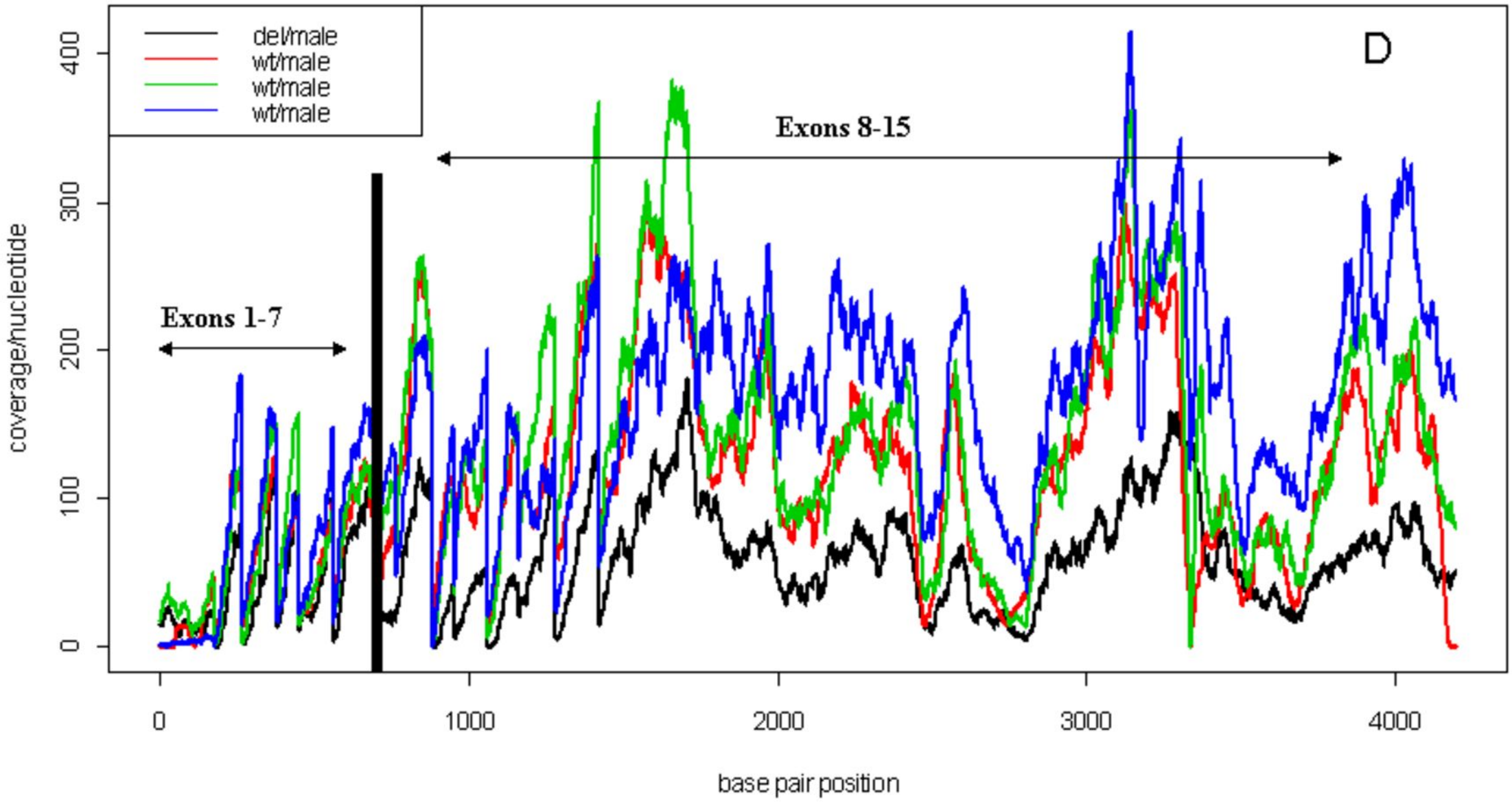

B3GALTL
C
Average coverage/exon
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
exon number
del/male
wt/male
wt/male
wt/male

Figure 2


del/male
wt/male
wt/male
wt/male
D
Exons 8-15
Exons 1-7
coverage/nucleotide
0
100
200
300
400
0
1000
2000
3000
4000
base pair position