Chapter 7:

Exon resequencing of genes within the bronx waltzer *candidate region*

CHAPTER 7

EXON RESEQUENCING OF GENES WITHIN THE *BRONX WALTZER* CANDIDATE REGION

7.1 INTRODUCTION

Following extensive mapping studies (see Chapters 3 and 4), the candidate region for *bronx waltzer* has proved difficult to refine and currently stands as a 2.6Mb region of chromosome 5 comprising 52 annotated genes. Since none of these have emerged as a clear causative agent for the phenotype, they must all be considered as potential candidates. This chapter describes the strategy utilised to carry out large-scale mutation screening in an effort to identify which of these genes is responsible for *bronx waltzer*, or alternatively to narrow down the list of candidates by demonstrating that they do not harbour any mutations in the *bronx waltzer* mutant genome.

7.1.1 Large scale mutation detection

Many strategies for large scale mutation detection rely on the differing properties of the two strands when a DNA sample is heterozygous. Differences in the secondary structure adopted by single-stranded molecules (Single Strand Conformation Polymorphism - SSCP) can be detected by their differing electrophoretic mobilities which allow samples differing by a single base pair to be separated by electrophoresis. Other methods rely on the differing properties of the heteroduplexes and homoduplexes that are formed during hybridisation of heterozygous molecules, with a heteroduplex generally having reduced electrophoretic mobility and a lower melting temperature. This latter means that samples can be analysed to high resolution using a specialized form of HPLC (High Performance Liquid Chromatography), and in the guise of WAVE analysis this has become a common method of detecting mutations in large numbers of samples in the fields of both research (Rossetti *et al.* 2002) and clinical diagnostics (Marsh *et al.* 2001; Ou-Yang *et al.* 2004).

7.1.1.1 WAVE analysis

The WAVE system developed by Transgenomic (Omaha, Nebraska, USA) operates by resolving heteroduplexes and homoduplexes and is based on the principle that heteroduplexes are less stable and therefore will be denatured at a lower temperature. Following PCR amplification the samples are denatured and then hybridised at gradually decreasing temperatures before being loaded onto a WAVE machine for the performance of HPLC. When performed at denaturing temperatures, this analysis results in a unique trace pattern which will differ depending on whether the sample contains only homoduplexes or a mixture of homoduplexes and heteroduplexes. Thus, where a mutation or polymorphism is present, the trace obtained for that sample will differ from the control homozygous trace, highlighting the amplicon for further analysis. An example of such data is shown in Figure 7.1. In the case of the *bronx waltzer* mutation, heterozygous samples would need to be compared to wild type or mutant samples in order to identify any differences. This can be easily achieved since in a normal maintenance mutant (*bv*/*bv*) by heterogygote (+/*bv*) mating, any animals which do not exhibit the bronx waltzer phenotype must have the genotype +/*bv*.

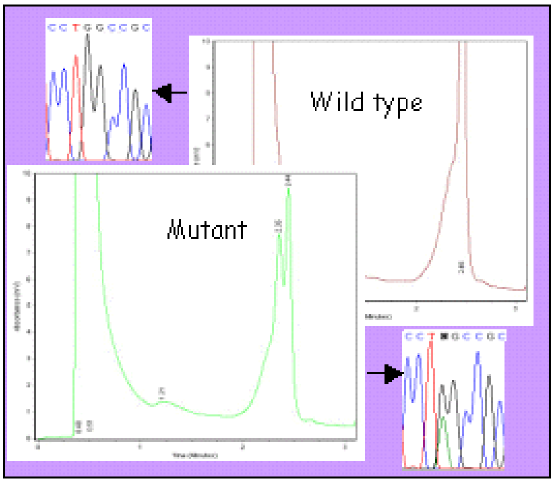


Figure 7.1: An example of the data obtained from WAVE analysis used for the detection of mutations. The mutant sample carries a shoulder on the second peak which is not present in the wild type, suggesting a variation in the two samples which was later confirmed by sequencing. Figure and data from the MRC Mammalian Genetics Unit, Harwell. (http://www.mgu.har.mrc.ac.uk/facilities/gems/mutdetect.html)

7.1.1.2 High-throughput exon resequencing

A straightforward method of determining whether a gene harbours a mutation in a given DNA sample is to obtain sequence reads for each of the exons and compare these to those obtained from a wild type control. This has the advantage that any differences are immediately identified without any further analysis required, and is also the most sensitive method of detecting mutations since it displays directly the alteration to the nucleotide sequence, rather than relying on any secondary effects of such a change which may vary in their intensity and thus in their chances of detection. In the light of this, and given the recent advances which make high-throughput sequencing a relatively routine procedure, it was decided that the genes in the *bronx waltzer* candidate region would be screened by exon resequencing.

Like other methods though, the success of this approach is limited by the regions of sequence that can be covered, which may be restricted by a number of factors. Firstly, exons can only be sequenced if they can first be amplified successfully by PCR, and even then they may prove difficult to sequence as a result of repeat regions, high GC content or a secondary structure which inhibits the progress of the DNA polymerase. Second, the resequencing of exons assumes that the causative mutation lies within a coding region or splice junction of a gene. While this is true in the majority of cases, it is possible that it may instead be located within a promoter or regulatory element, the exact position and features of which are not easily defined, and thus it would be difficult to include these in the screen. Finally, exons can only be resequenced if their location is known, and thus the coverage obtained is limited by the quality of published sequence and gene annotation within the candidate region. This point is discussed in the following section.

7.1.2 The mouse genome sequence

The published sequence for the mouse is derived from the strain C57BL/6J and was compiled by composite assembly of whole genome shotgun (WGS) and High Throughput Genome Sequence (HTGS) sequence data (Waterston *et al.* 2002). End sequencing and fingerprinting of BAC libraries allowed the creation of a tiling path for each chromosome (Gregory *et al.* 2002) around which the sequence reads were arranged to give the NCBI Mouse Builds 30-33. The currently available Build 33 comprises HTGS sequence from clones which are either finished or in the advanced stages of sequencing, with WGS sequence being used to bridge any gaps and give an overall relatively high degree of sequence coverage. However, this stage of the project is currently underway, meaning that different regions of the mouse genome presently have varying levels of sequence quality, as illustrated in Figure 7.2.

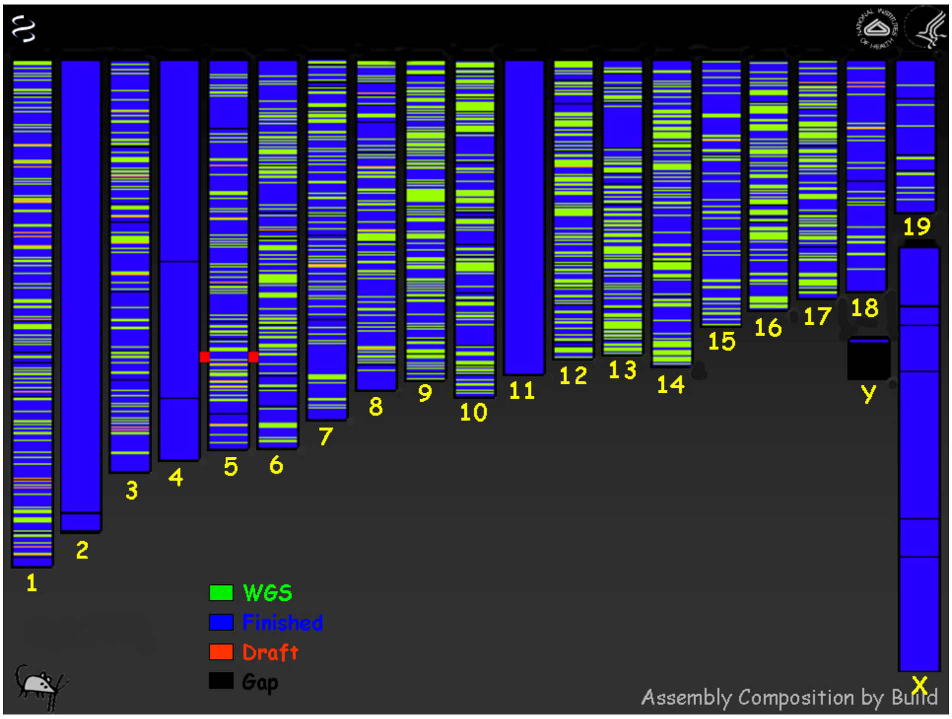


Figure 7.2: Graphical representation of the sequencing status across the mouse genome. The location of the *bronx waltzer* candidate region is highlighted by red bars. Diagram adapted from NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html).

7.1.2.1 Sequence quality in the *bv* candidate region

The *bronx waltzer* candidate region spans from 111609074 to 114063755bp of chromosome 5 in Ensembl Build33, with the whole region being contained within the single mapping contig 5015b (see Figure 7.3). This contiguous mapping coverage allows it to be said with reasonable certainty that the size and ordering of the region as a whole has been well established. Fifteen BACs from this map have been selected for HTGS sequencing with four of these being finished and all the others having some sequence available (See Table 7.1). Two gaps exist in the BAC sequence tiling path at present (see Figure 7.3), which are partially covered by WGS sequence. In sequencing terms, the region is currently spanned with a total of five sequence gaps, with three of these coinciding with one of the tiling path gaps. This is not a region where BAC sequencing has been prioritised as yet, although discussions are underway in order to facilitate this.

Accession Number	**Clone**	**Sequencing Centre**	**Sequence Status**
AC122282	RP23-240B12	WUGSC	finish_start
AC127255	RP24-291I7	WUGSC	submitted
AC132102	RP24-360K9	WUGSC	submitted
AC124228	RP23-345O14	WIBR	working draft sequence
AC087330	RP23-3M10	HPGC	complete sequence
AC116500	RP24-297N9	WIBR	sequencing in progress
AC117799	RP24-556H22	WIBR	working draft sequence
AC119205	RP24-188M1	WIBR	working draft sequence
AC145070	RP23-41I11	WIBR	complete sequence
AC117735	RP24-132L16	WIBR	working draft sequence
GAP			
AC139562	RP24-180G11	WIBR	sequencing in progress
AC113200	RP23-265I19	WIBR	working draft sequence
AC115795	RP23-400O10	WIBR	working draft sequence
AC119977	RP24-486F14	WIBR	sequencing in progress
GAP			
AC115972	RP24-79A6	WIBR	sequencing in progress

Table 7.1: Sequence status of BACs selected for High Throughput Genome Sequencing (HTGS) across the *bronx waltzer* candidate region.
Those clones with a status of "finish_start", "complete sequence" or "submitted" are in the more advanced stages of sequencing and generally have 5-10 fold coverage along their length. Those with a status of "working draft sequence" or "sequencing in progress" are in the earlier stages of sequencing and generally have 3-6 fold coverage.

Abbreviations of sequencing centres:
WUGSC – Washington University genome Sequencing Center
WIBR – Whitehead Institute for Biomedical Research (BROAD Institute)
HPGC – Harvard Partners Genome Center

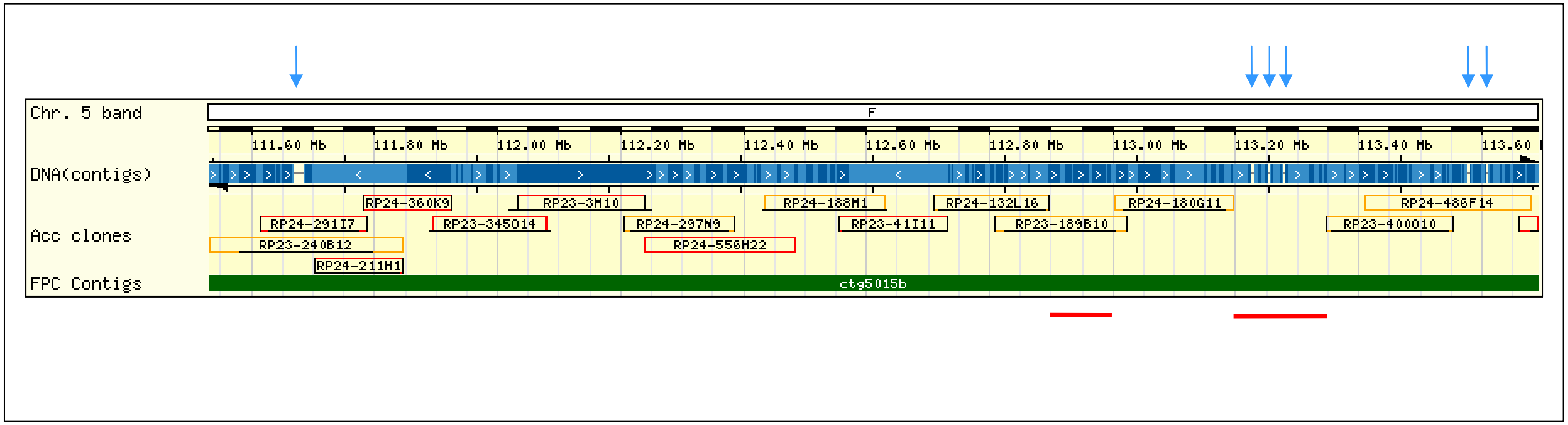


Figure 7.3: Map showing sequence and BAC coverage across the *bronx waltzer* candidate region, taken from Ensembl CytoView (http://www.ensembl.org).
Gaps in the DNA (contigs) represent areas where no sequence is available and are highlighted by **blue arrows**.
Accessioned clones are shown as red and yellow boxes, and gaps in the tiling path are highlighted by **red bars**. The clone RP23-189B10 has been "abandoned", and therefore does not represent part of the tiling path.
The region is contained within one single mapping contig – ctg5015b – suggesting that the gross structure and order of the interval has been well established.

7.1.2.2 Gene annotation

Gene annotation of the mouse genome has been based on a number of sources of information. Some genes were already well characterised as a result of earlier studies and could thus be immediately annotated. In other cases, the homology to EST and cDNA sequences or genes in other species can provide strong evidence for the annotation of a gene, although genes which are expressed at low levels or in very specific tissues or periods of development may be underrepresented when such an approach is used. A powerful tool in the annotation of genomes has been the development of gene prediction algorithms based on hidden Markov models (HMMs) and related models such as Genscan (Burge and Karlin 1997), GeneMark (Borodovsky and McIninch 1993) and FGENE (Solovyev *et al.* 1995). While none of these is foolproof, taken together they can provide analysis of potential splice sites, 5'-coding, internal exon, and 3'-coding regions, as well as taking into account GC content, CpG islands, promoter sites and open reading frames, amongst other indicators. This can provide additional evidence in the case of existing EST sequence, or can act as a hypothesis in a directed search for corroborating experimental evidence via systematic RT-PCR and sequencing. Historically, this type of analysis has lead to the over-prediction of genes as a result of the presence of pseudogenes within genomes. The development of dual-genome *de novo* predictors such as SLAM (Alexandersson *et al.* 2003) and TWINSCAN (Korf *et al.* 2001; Flicek *et al.* 2003) has helped to eliminate many such false positives (Brent and Guigo 2004)

7.2 METHODS

7.2.1 Template DNA samples

Exons were amplified using template DNA from a *bronx waltzer* mutant mouse (*bv*/*bv*) and a wild type control (+/+) mouse, both from the original *bronx waltzer* colony. Since the background on which *bv* is maintained is unknown, it was important to compare the mutant sequence to equivalent sequence from a wild type mouse on the same background rather than simply comparing it to the published sequence in order to eliminate the confusion caused by SNPs and other non-pathogenic variations between strains.

Since *bronx waltzer* is recessive, heterozygote (+/*bv*) and wild type (+/+) mice in the *bronx waltzer* stock are phenotypically indistinguishable. Therefore, as no genotyping tool exists for this mutation, test matings were established in order to identify a mouse carrying two unaffected *bv* alleles. In the first stage, two heterozygote mice were paired in order to give offspring which could be typed as either *bv*/*bv* if they manifested the *bronx waltzer* phenotype or as +/? if they did not. Second, +/? mice were paired with a male homozygous *bv*/*bv* mouse in a test mating to verify their genotype. If any of the offspring showed a *bronx waltzer* phenotype then they must have inherited a mutated allele from each parent, thus confirming that the uinknown genotype must be +/*bv*. Should all the offspring appear normal then they must all be inheriting a wild type allele from the unknown parent. After 21 classifiable births (offspring which survive and can be phenotyped) which all show no signs of the *bronx waltzer* phenotype, the genotype can be confirmed as +/+. This strategy is illustrated in Figure 7.4.

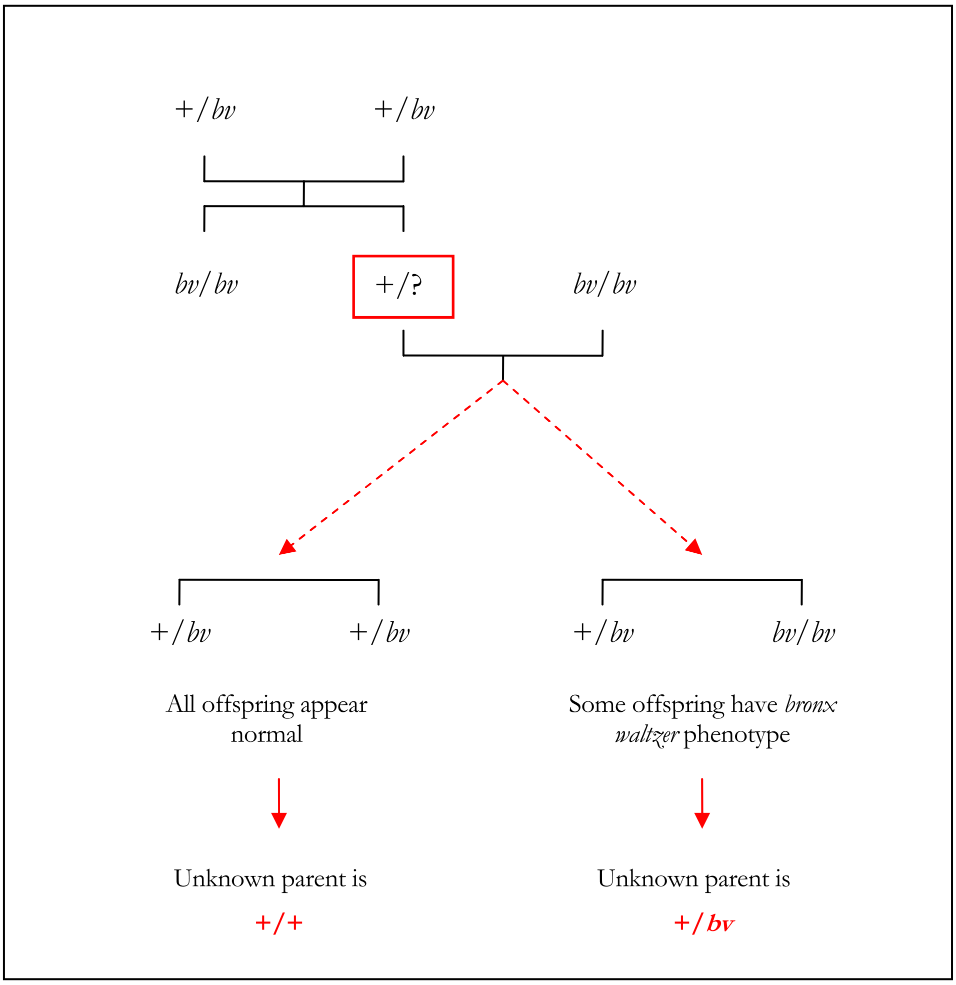


Figure 7.4: Mating strategy employed in the identification of a mouse from the *bronx waltzer* colony which could be confirmed as carrying two wild type copies of the *bv* allele.
Mice from a mating of two known *bronx waltzer* heterozygotes (+/*bv*) which do not manifest a phenotype may be either +/+ or +/*bv* and are designated +/? (red box). In order to determine their genotype, they can be mated with a *bronx waltzer* homozygous mutant, with the phenotype of their offspring revealing whether or not they carry a mutant *bv* allele.

In the case of *bronx waltzer*, three +/? females were paired with *bv*/*bv* males in order that the male could be removed from the cage before birth to avoid the problem of offspring being cannibalised by a hyperactive parent. Of these three females, two gave birth to offspring which were typed as *bronx waltzer* and thus their genotypes were confirmed as +/*bv*. One female gave birth to 24 offspring, of which 19 were classifiable (the remainder dying prior to being old enough to determine their phenotype) and all of which were normal in behaviour. Although this was two short of the 21 classifiable births required in order to definitively designate the mouse as +/+, the mating pair had stopped producing offspring and so tissue was taken for DNA extraction in the relative certainty that the genotype was +/+.

7.2.2 Large-scale primer design

The large number of amplicons required to obtain sequence for every gene in the *bronx waltzer* candidate region necessitated the employment of a semi-automated system for primer design. This was achieved using the primer design pipeline established at the Sanger Institute, which uses Ensembl Gene IDs to retrieve the flanking sequences of annotated exons. These are then processed in a similar manner to that employed by Primer3 (Rozen and Skaletsky 2000) with potential primers being screened for similarity to repeat sequences and for self-complementarity. All the primers used correspond to intron sequence such that coding sequence and intron/exon boundaries were examined for mutations.See Appendix C for primer sequences.

7.2.3 Nested PCR amplification

Those exons for which good quality sequence was not obtained in the first pass of amplification were then amplified using nested PCR. This method employs two rounds of PCR, with the second set of PCR primers lying within the fragment amplified by the first. Since these will only amplify from the desired sequence and not from any non-specific products generated in the first round, the result is a higher quantity of a more specific product which in turn can be used as template DNA to give higher quality sequence traces with reduced background. This strategy is illustrated in Figure 7.5

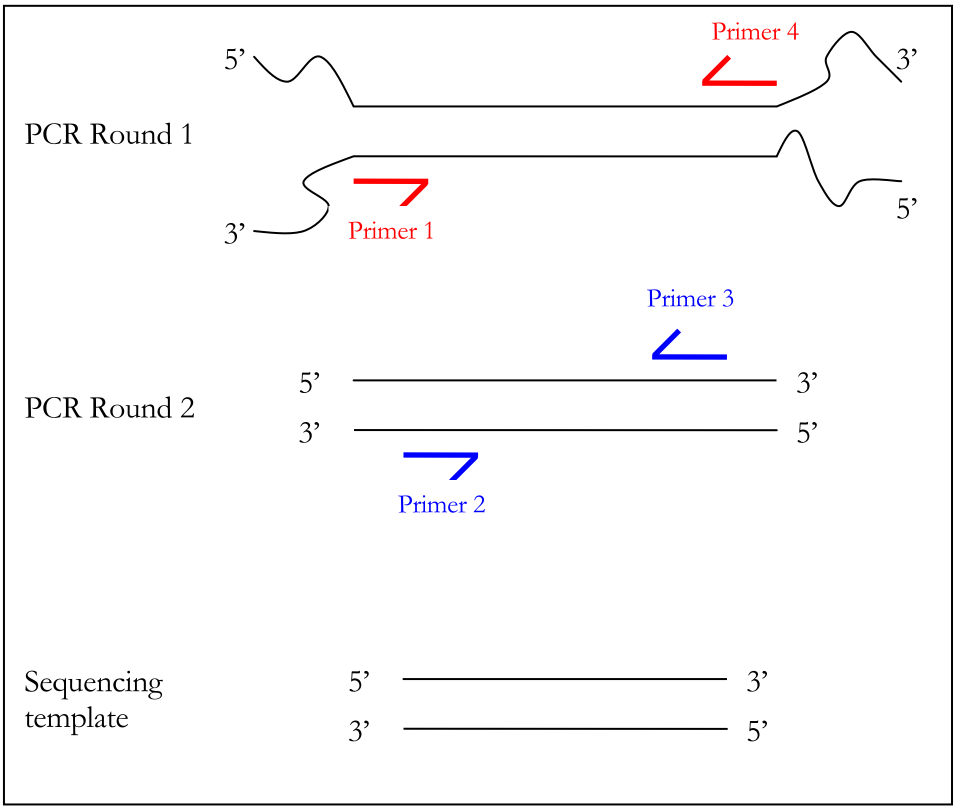


Figure 7.5: Nested PCR allows the elimination of non-specific products sometimes generated in single stage PCR which can give rise to poor quality sequence reads. The second round of PCR uses the product from the first round as template DNA, and the primers are designed to bind to sequence within the desired amplicon.

The design of nested PCR primers was carried out using LIMSTILL (LIMS for Identification of Mutations by Sequencing and TILLING; http://limstill.niob.knaw.nl/), an open-source software developed by the Hubrecht Laboratory (Netherlands). This allows querying of the Ensembl database for the selection of exons for amplification, followed by primer design using a Perl/CGI interface. Primer sequences are given in Appendix C.

The first round of PCR was carried out using the standard PCR protocol described in Section 2.4. In the second round of PCR, the concentration of primers was limited in order to generate final products of roughly equal concentrations, allowing the subsequent sequencing reactions to be performed at optimal concentration and thus giving higher quality sequence. In this study, primers were diluted to a concentration of between 100 – 130nM to give an approximate amplicon concentration of 115nM.

7.2.4 Sequence analysis

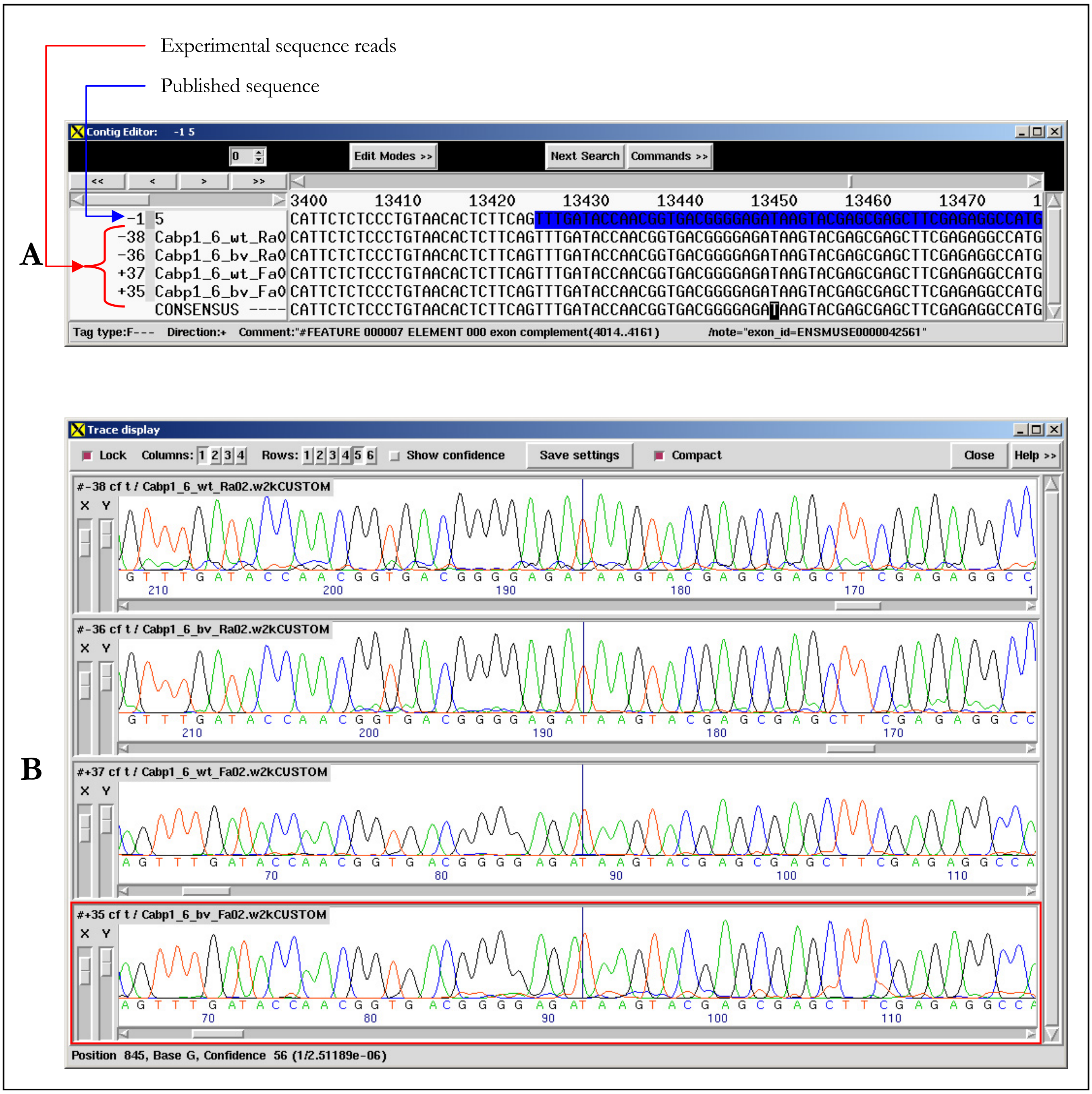
DNA sequencing was carried out as described in Section 2.8 and the sequence reads analysed using the software package Gap4 (Bonfield *et al.* 1998). Analysis consisted of the creation of a database containing all the sequence reads for a given gene alongside a reference flat file exported from Ensembl which included the annotated positions of exons. These sequences were then assembled into contigs and examined for differences between those originating from a *bv* mutant template and those amplified from the wild type control. A list of every exon within the candidate region was created and annotated with details of the sequence coverage obtained. Coverage of each exon in both DNA samples was preferred, but where a region was not covered by wild type sequence and the mutant sequence matched that of the published strain, the exon was considered to be covered. Where partial sequence of an exon was obtained, the distance remaining was recorded and the percentage of base pairs covered for each gene was calculated.

7.3 RESULTS

7.3.1 Sequence coverage

A list was compiled of all the annotated exons within the *bv* candidate region. This spans the distance between the flanking markers DASNP3 and D5Mit209 and extends from 111609074bp to 114063755bp of chromosome 5 in Mouse Ensembl Build33. Where more than one alternative transcript was given for a particular gene, each unique exon was only included once with the largest being selected where they differed in size. Coding sizes of each exon were obtained, with untranslated exons assigned a value of zero, and these were used to calculate the figure for the total number of coding base pairs within the region. At the time of writing, the region contains 534 exons which constitute 52 genes and include 72523 base pairs of coding sequence.

Following the first round of primer design and sequencing, approximately 72% of the coding exons had good sequence coverage, but only four genes were fully sequenced. It was thus necessary to carry out a second round using the modifications describe in section 7.2.3 to improve the quality of the data. Following this step, 91% of coding exons had good coverage and 22 genes were fully sequenced. The total number of base pairs covered using both sets of data was 65608bp, representing 90.5% of the total coding sequence within the region. Examples of the sequence obtained by both approaches are given in Figure 7.6, and the sequence coverage obtained for each gene within the candidate region is represented in Figure 7.7.



Contig Editor: -1 5

```
                          3400      13410     13420     13430     13440     13450     13460     13470     1
 -1  5                    CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
-38  Cabp1_6_wt_Ra0       CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
-36  Cabp1_6_bv_Ra0       CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
+37  Cabp1_6_wt_Fa0       CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
+35  Cabp1_6_bv_Fa0       CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
     CONSENSUS ----       CATTCTCTCCCTGTAACACTCTTCAGTTTGATACCAACGGTGACGGGGAGATAAGTACGAGCGAGCTTCGAGAGGCCATG
```

Tag type:F--- Direction:+ Comment:"#FEATURE 000007 ELEMENT 000 exon complement(4014..4161) /note="exon_id=ENSMUSE0000042561"

Trace display

#-38 cf t / Cabp1_6_wt_Ra02.w2kCUSTOM

X Y
G T T T G A T A C C A A C G G T G A C G G G G A G A T A A G T A C G A G C G A G C T T C G A G A G G C C
210 200 190 180 170 1

#-36 cf t / Cabp1_6_bv_Ra02.w2kCUSTOM

X Y
G T T T G A T A C C A A C G G T G A C G G G G A G A T A A G T A C G A G C G A G C T T C G A G A G G C C
210 200 190 180 170

#+37 cf t / Cabp1_6_wt_Fa02.w2kCUSTOM

X Y
A G T T T G A T A C C A A C G G T G A C G G G G A G A T A A G T A C G A G C G A G C T T C G A G A G G C C A
70 80 90 100 110

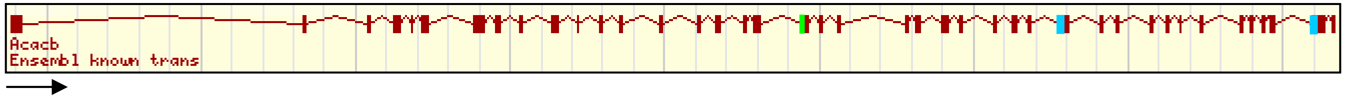
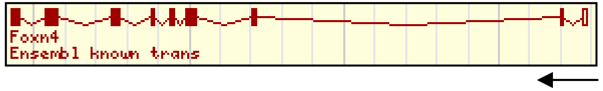
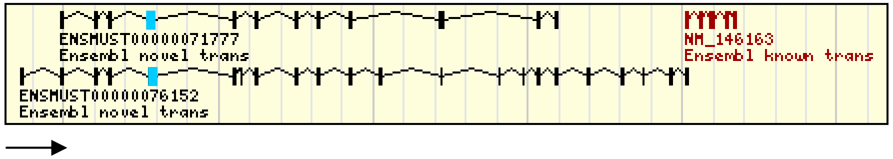
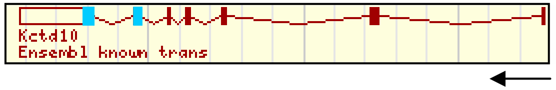
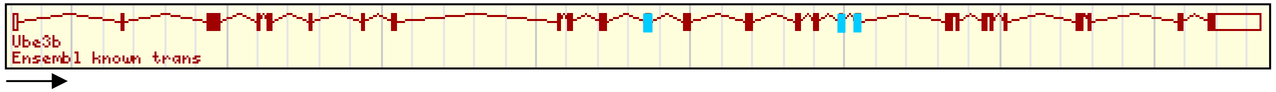
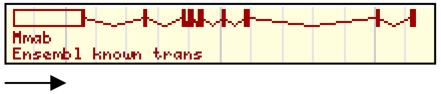
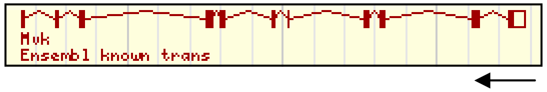
#+35 cf t / Cabp1_6_bv_Fa02.w2kCUSTOM

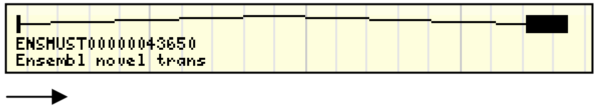
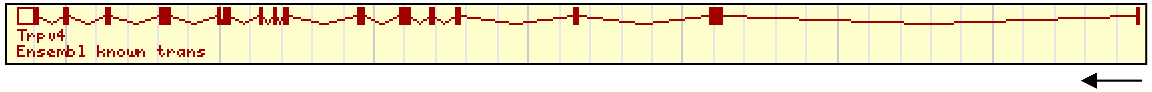
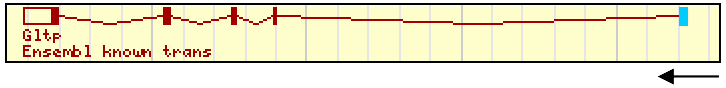
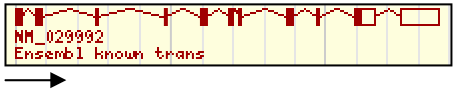
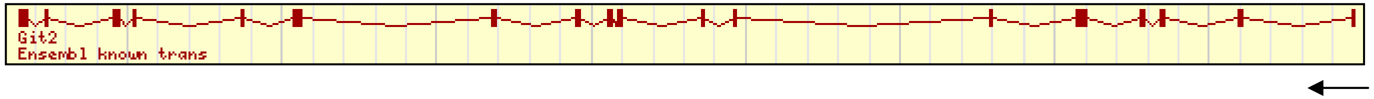
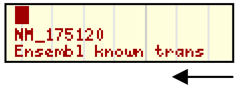
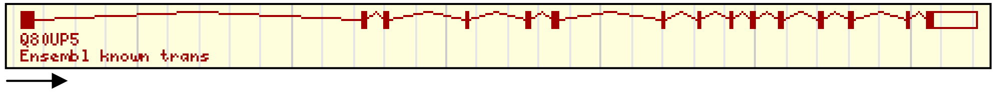
X Y
A G T T T G A T A C C A A C G G T G A C G G G G A G A T A A G T A C G A G C G A G C T T C G A G A G G C C A
70 80 90 100 110

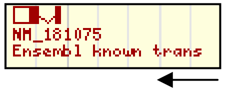
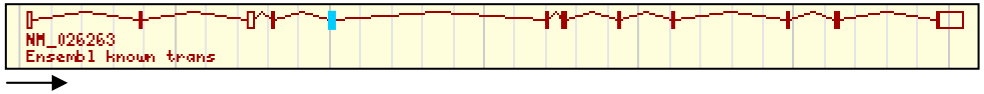
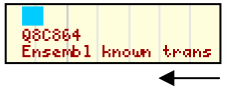
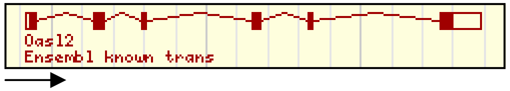
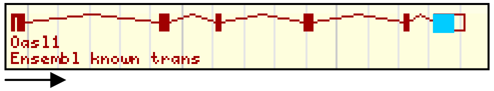
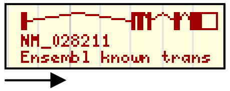
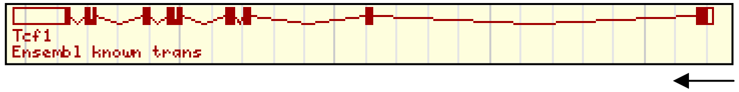
Position 845, Base G, Confidence 56 (1/2.51189e-06)

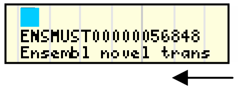
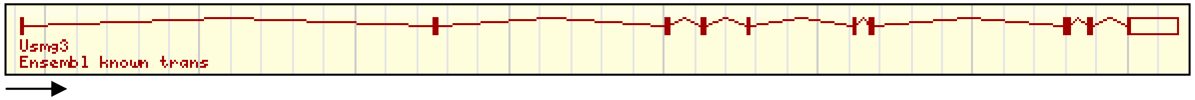
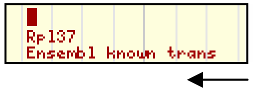
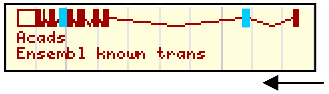
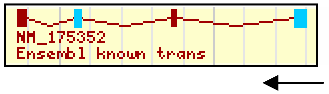
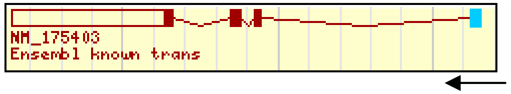
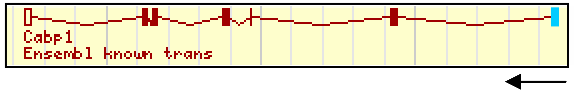
Figure 7.6: Screenshots from the software package gap4, illustrating the manner in which sequence reads were analysed. In the Contig Editor (A), the forward and reverse sequence reads obtained from wild type and mutant templates are aligned to each other and to the published sequence for comparison. Exons annotated onto the published sequence are highlighted in blue and the exon identifier is shown in the data bar at the bottom of the screen, making assessment of exon coverage relatively simple. Any discrepancies between the sequence reads are also automatically highlighted any which occur can be examined by visualising the sequence reads in the Trace Display window (B).

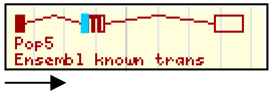
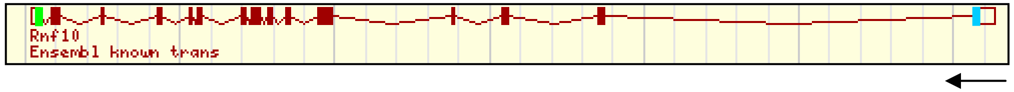
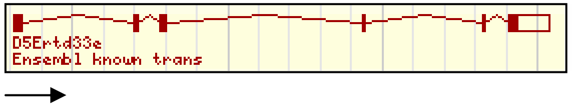
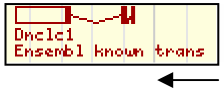
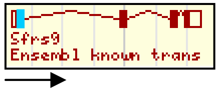
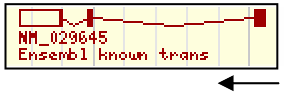
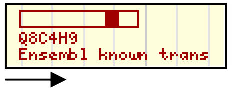
Figure 7.7: (Overleaf)
Sequence coverage obtained for genes within the *bronx waltzer* candidate region.
Exons are represented as coloured boxes, with those for which *bv* mutant sequence has been obtained and analysed shown in **DARK RED** (for known genes) or **BLACK** (for predicted genes). Exons for which partial sequence is available are shown in **GREEN**, and those which are unsequenced are represented in **BLUE**.
The diagrams and figures represent coding sequence only. Sequencing of untranslated regions (**YELLOW**) was attempted but the data is not shown.
The direction of transcription is shown by a black arrow, with those reading left to right being annotated on the plus strand and those reading left to right on the minus strand.
Genes are given in the order in which they are annotated along chromosome 5, from the centromeric to the telomeric end of the region. Graphical representations of the genes are taken from Ensembl Build33.
Genes marked by an asterisk were not annotated at the time of primer design, and therefore no sequencing of these has been attempted.

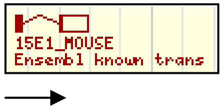
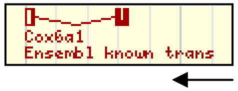
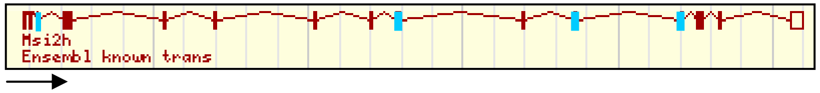
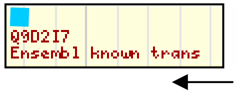
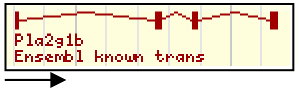
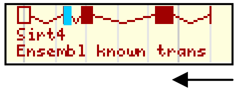
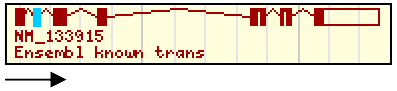
Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
Acetyl-coenzyme A carboxylase beta  	7373	7136	96.8%
Forkhead box protein N4  	1557	1557	100%
Myosin 1h  ENSMUST00000071777 Ensembl novel trans ENSMUST00000076152 Ensembl novel trans NM_146163 Ensembl known trans	2846	2766	97.2%
Potassium channel tetramerization domain-containing 10  	933	623	66.8%
Ubiquitin protein ligase E3B  	3187	2797	87.8%
Methylmalonic aciduria type B homolog  	735	735	100%
Melavonate kinase  	1178	1178	100%

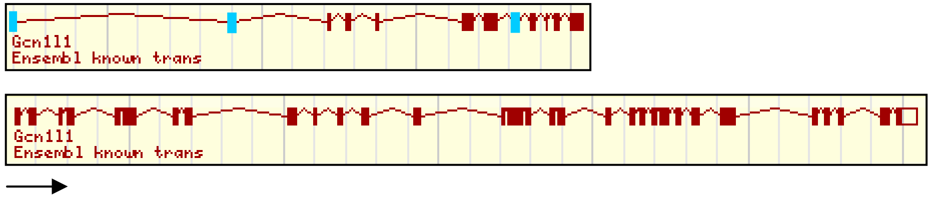
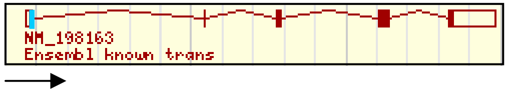
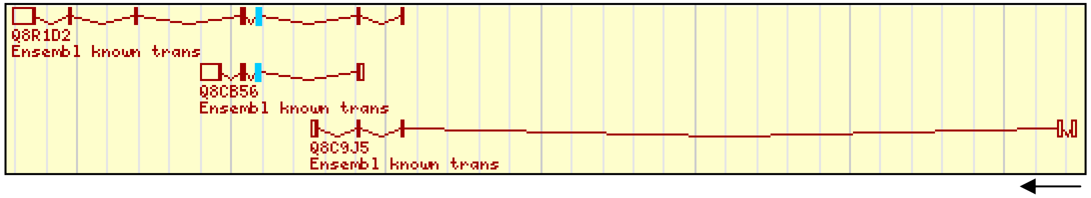
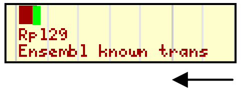
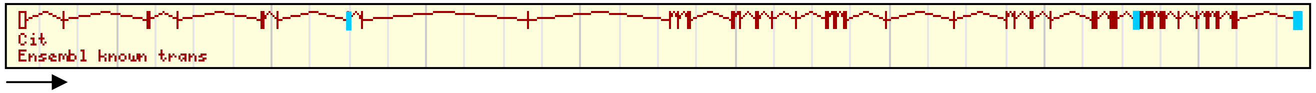
Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
ENSMUSG00000041930  	1362	1362	100%
Transient receptor potential cation channel V4  	2601	2601	100%
Glycolipid transfer protein  	625	523	83.7%
NM_029992  	1379	1379	100%
G protein-coupled receptor kinase-interactor 2  	2126	2126	100%
NM_175120  	374	374	100%
Q9D1S6  	2031	2031	100%

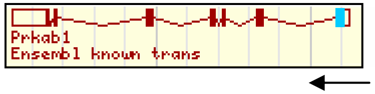
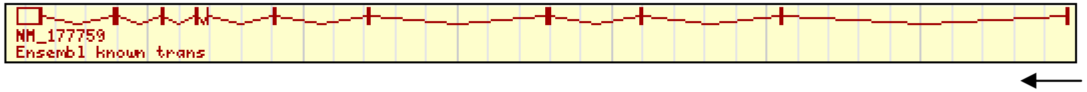
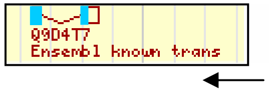
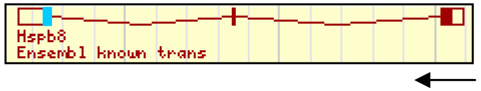
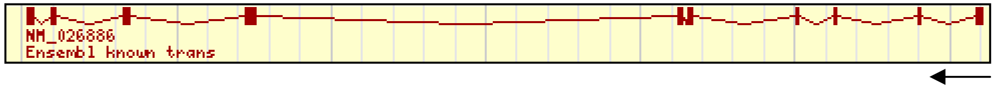
Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
NM_181075  	418	418	100%
NM_026263  	547	463	84.6%
Q8C864 *  	482	0	0%
2'-5'-oligoadenylate synthetase like protein 2  	1398	1398	100%
2'-5' oligoadenylate synthetase-like 1  	1530	1030	67.3%
NM_028211  	765	765	100%
Hepatocyte nuclear factor 1-alpha  	1877	1877	100%

Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
ENSMUSG00000044804 *  	378	0	0%
Upregulated during skeletal muscle growth 3  	1345	1345	100%
60S ribosomal protein L37  	293	293	100%
Acyl-CoA dehydrogenase, short-chain specific  	1229	929	75.6%
NM_175352  	712	362	50.8%
NM_175403  	871	640	73.5%
Calcium-binding protein 1  	680	636	93.5%

Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
Processing of precursor 5, ribonuclease P  	516	379	73.4%
Ring finger protein 10  	2398	2255	94%
D5Ertd33e  	977	977	100%
Dynein light chain 1, cytoplasmic  	268	268	100%
Splicing factor, arginine/serine rich 9  	665	475	71.4%
NM_029645  	465	465	100%
Q8C4H9 *  	341	341	0%

Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
15E1_MOUSE 	229	229	100%
Cytochrome c oxidase polypeptide VIa-liver 	336	336	100%
Musashi homolog 1 	1016	736	72.4%
Q9D2I7 * 	464	0	0%
Phospholipase A2 precursor 	437	437	100%
NAD-dependent deacetylase sirtuin 4 	990	781	78.9%
Paxillin alpha 	1787	1672	93.6%

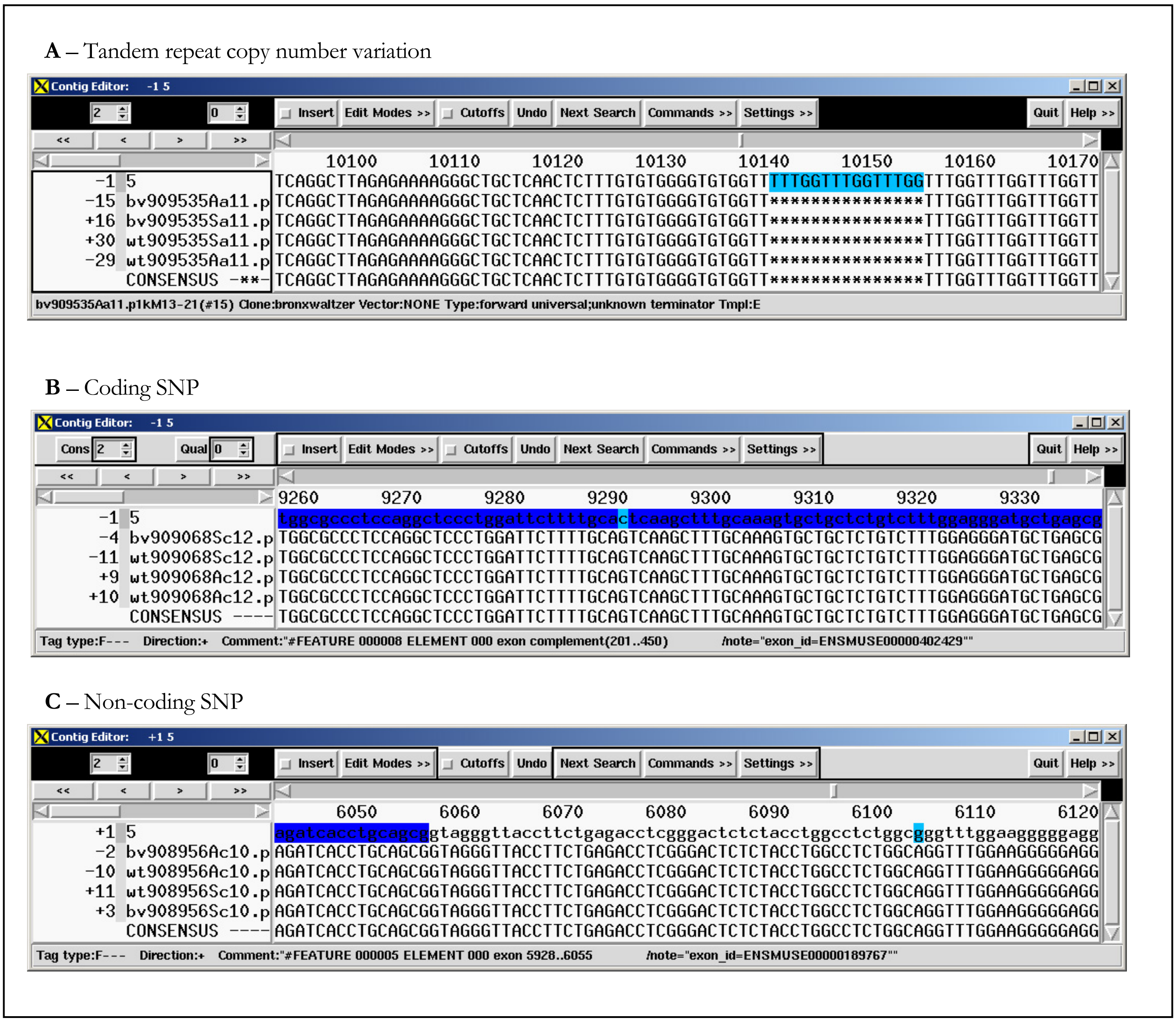
Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
General control of amino-acid synthesis 1-like 1  Gcn1l1 Ensembl known trans Gcn1l1 Ensembl known trans	6752	6614	98%
NM_198163  	600	549	91.5%
Q8R1D2  Q8R1D2 Ensembl known trans Q8CB56 Ensembl known trans Q8C9J5 Ensembl known trans	2239	2109	94.2%
60S ribosomal protein L29  	482	392	81.3%
Citron protein  	6220	5859	94.2%

Gene name and structure	Coding sequence (bp)	Sequence covered (bp)	% coverage
5'-AMP-activated protein kinase, beta-1 subunit 	808	648	80.4%
NM_177759 	1432	1432	100%
Q9D4T7 * 	340	0	0%
Heat-shock protein beta-8 	588	429	72.9%
NM_026886 	1343	1343	100%

7.3.2 Sequence polymorphisms

During the analysis of sequence data, several polymorphisms were identified. In several locations, polymorphisms were found where the *bronx waltzer* mutant and wild type sequences matched each other but differed from the published sequence. These variations included coding and non-coding SNPs, as well as a number of tandem repeat copy number variations, examples of these are shown in Figure 7.8.

In addition, in a localised portion of the candidate region centred on the gene Citron, a number of polymorphisms were identified where the *bronx waltzer* and wild type sequences differed from each other. These are described and illustrated in Table 7.2. All of these coding nucleotide substitutions were silent, resulting in no change to the peptide sequence.



A – Tandem repeat copy number variation

```
                    10100      10110      10120      10130      10140      10150      10160      10170
 -1 5               TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTTTTTGGTTTGGTTGGTTTGGTTTGGTTTGGTT
-15 bv909535Aa11.p  TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTT**************TTTGGTTTGGTTTGGTT
+16 bv909535Sa11.p  TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTT**************TTTGGTTTGGTTTGGTT
+30 wt909535Sa11.p  TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTT**************TTTGGTTTGGTTTGGTT
-29 wt909535Aa11.p  TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTT**************TTTGGTTTGGTTTGGTT
    CONSENSUS -**-  TCAGGCTTAGAGAAAAGGGCTGCTCAACTCTTTGTGTGGGGTGTGGTT**************TTTGGTTTGGTTTGGTT
bv909535Aa11.p1kM13-21(#15) Clone:bronxwaltzer Vector:NONE Type:forward universal;unknown terminator Tmpl:E
```

B – Coding SNP

```
                    9260      9270      9280      9290      9300      9310      9320      9330
 -1 5               tggcgccctccaggctccctggattcttttgcactcaagctttgcaaagtgctgctctgtctttggagggatgctgagcg
 -4 bv909068Sc12.p  TGGCGCCCTCCAGGCTCCCTGGATTCTTTTGCAGTCAAGCTTTGCAAAGTGCTGCTCTGTCTTTGGAGGGATGCTGAGCG
-11 wt909068Sc12.p  TGGCGCCCTCCAGGCTCCCTGGATTCTTTTGCAGTCAAGCTTTGCAAAGTGCTGCTCTGTCTTTGGAGGGATGCTGAGCG
 +9 wt909068Ac12.p  TGGCGCCCTCCAGGCTCCCTGGATTCTTTTGCAGTCAAGCTTTGCAAAGTGCTGCTCTGTCTTTGGAGGGATGCTGAGCG
+10 wt909068Ac12.p  TGGCGCCCTCCAGGCTCCCTGGATTCTTTTGCAGTCAAGCTTTGCAAAGTGCTGCTCTGTCTTTGGAGGGATGCTGAGCG
    CONSENSUS ----  TGGCGCCCTCCAGGCTCCCTGGATTCTTTTGCAGTCAAGCTTTGCAAAGTGCTGCTCTGTCTTTGGAGGGATGCTGAGCG
Tag type:F---   Direction:+   Comment:"#FEATURE 000008 ELEMENT 000 exon complement(201..450)     /note="exon_id=ENSMUSE00000402429""
```

C – Non-coding SNP

```
                        6050      6060      6070      6080      6090      6100      6110      6120
 +1 5               agatcacctgcagcggtagggttaccttctgagacctcgggactctctacctggcctctggcgggtttggaagggggagg
 -2 bv908956Ac10.p  AGATCACCTGCAGCGGTAGGGTTACCTTCTGAGACCTCGGGACTCTCTACCTGGCCTCTGGCAGGTTTGGAAGGGGGAGG
-10 wt908956Ac10.p  AGATCACCTGCAGCGGTAGGGTTACCTTCTGAGACCTCGGGACTCTCTACCTGGCCTCTGGCAGGTTTGGAAGGGGGAGG
+11 wt908956Sc10.p  AGATCACCTGCAGCGGTAGGGTTACCTTCTGAGACCTCGGGACTCTCTACCTGGCCTCTGGCAGGTTTGGAAGGGGGAGG
 +3 bv908956Sc10.p  AGATCACCTGCAGCGGTAGGGTTACCTTCTGAGACCTCGGGACTCTCTACCTGGCCTCTGGCAGGTTTGGAAGGGGGAGG
    CONSENSUS ----  AGATCACCTGCAGCGGTAGGGTTACCTTCTGAGACCTCGGGACTCTCTACCTGGCCTCTGGCAGGTTTGGAAGGGGGAGG
Tag type:F---   Direction:+   Comment:"#FEATURE 000005 ELEMENT 000 exon 5928..6055     /note="exon_id=ENSMUSE00000189767""
```

Figure 7.8: Examples of polymorphisms identified between the *bv* genetic background and the published sequence. In each case the published sequence is shown as the top line and assigned the identifier "5". Exonic sequence is highlighted in **ROYAL BLUE**, while discrepancies between the sequences are highlighted in **SKY BLUE**.

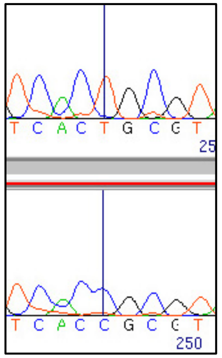
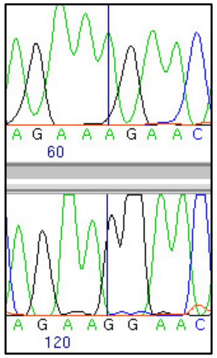
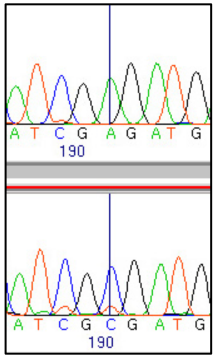
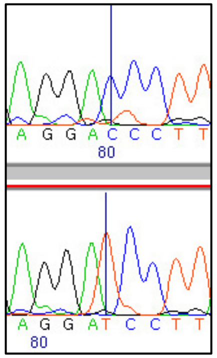
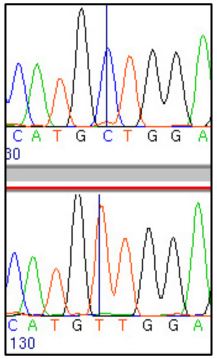
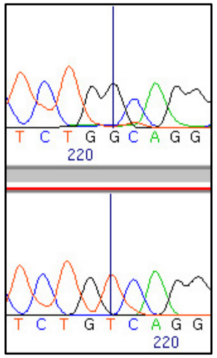
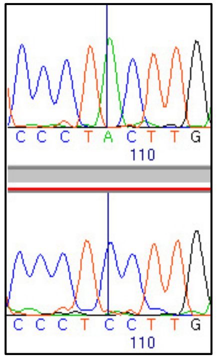
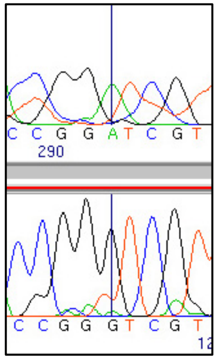
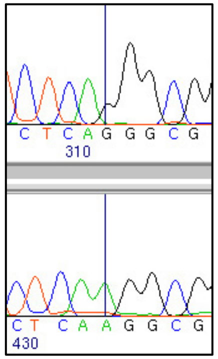
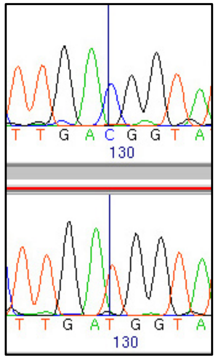
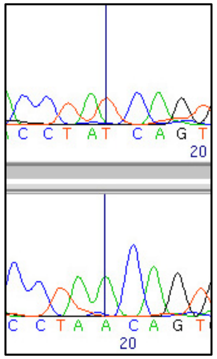
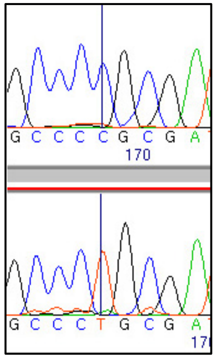
Polymorphism	1	2	3	4	5	6
Exon	ENSMUSE00000464783	ENSMUSE00000462386	ENSMUSE00000265597	ENSMUSE00000265591	ENSMUSE00000499740	ENSMUSE00000265576
Chromosomal location	113364973	113367988	113388242	113395210	113396084	113398326
Location relative to exon	Downstream	Coding	Coding	Downstream	Coding	Upstream
bv allele	T	A	A	C	C	G
wt allele	C	G	C	T	T	T
Sequence traces *bv* wt						
Codon change	-	AAA → AAG	CGA → CGC	-	CTG → TTG	-
Amino acid change	-	K → K	R → R	-	L → L	-
Consequence	-	Silent	Silent	-	Silent	-

Table 7.2: Locations, alleles and effects of SNPs identified between the bv/bv sequence and +/+ sequence from the same genetic background (continued overleaf). All of these polymorphisms are located within either introns or exons of the gene Citron.

Polymorphism	**7**	**8**	**9**	**10**	**11**	**12**
Exon	ENSMUSE00000265562	ENSMUSE00000265546	ENSMUSE00000265528	ENSMUSE00000264882	ENSMUSE00000264868	ENSMUSE00000264854
Chromosomal location	113406661	113409285	113412602	113421086	113422608	113424716
Location relative to exon	Upstream	Downstream	Downstream	Coding	Downstream	Coding
***bv* allele**	A	A	G	C	T	C
wt allele	C	G	A	T	A	T
Sequence traces ***bv*** **wt**						
Codon change	-	-	-	GAC → GAT	-	CCC → CCT
Amino acid change	-	-	-	D → D	-	P → P
Consequence	-	-	-	Silent	-	Silent

7.4 DISCUSSION

7.4.1 Sequence coverage

Following two rounds of primer design and sequencing, a total of 65729 base pairs of coding sequence, representing 90.63% of that lying within the *bv* candidate region have been successfully sequenced in the *bronx waltzer* mutant mouse genome. The use of nested PCR primers in the second round allowed a significant improvement in the level of sequence coverage, since the resulting sequence reads were longer and more consistent. In order to achieve full sequence coverage a number of further steps could be taken. Initially, the PCR conditions for the existing primer sets could be altered in an attempt to amplify a product of high enough concentration and specificity for sequencing. Initial adjustments could include alteration of the annealing temperature and magnesium concentration. In the case of fragments in regions with high GC content or secondary structure, modifications such as increasing the denaturation temperature, the addition of DMSO and betaine, or linearisation of the template can help to prevent polymerase slippage. There are many GC-rich PCR kits available which perform these functions, as well as including a proofreading enzyme to maximise the chances of obtaining a specific PCR product. Exons which still fail primer design or amplification from genomic DNA could then be amplified instead from a cDNA template using primers designed in adjacent exons. This would circumvent any problematic intronic repeat regions which cause difficulties in amplification and sequencing, or regions where no sequence is available and therefore primers cannot be designed. Finally, exons for which a PCR product can be obtained but which prove difficult to sequence may be subcloned into a vector in order to minimize the amount of high GC template and reduce the opportunity for stable secondary structures to form. In addition, T7 and SP6 primers can then be used in the sequencing reactions, thus eliminating any issues with the suitability of the PCR primers for use in sequencing.

7.4.2 Candidacy of genes following re-sequencing

No mutations have so far been discovered in the *bv* candidate region, and thus the gene responsible for the mutation has not been identified, although the coding regions of 23 genes have been fully sequenced in the mutant genome, although it is not possible to completely rule out the involvement of these genes since a pathogenic mutation may lie in a regulatory element (see Section 7.1.1.2). Indeed, given the highly specific nature of the *bronx waltzer* phenotype, it is plausible that the mutation may lie in a regulatory element controlling expression of the gene in the cell types affected – that is the inner hair cells of the organ of Corti and the hair cells of the vestibular system. The implications of this are discussed further in Section 9.1.3. Even so, the successful resequencing of the complete coding region of a gene can be said to significantly reduce the probability of its being the causative agent for the *bronx waltzer* phenotype. Those genes for which full coding sequence has been obtained are listed below in Table 7.3.

Gene symbol	Description
Foxn4	Forkhead box protein N4
Mmab	Methylmalonic aciduria type B homolog
Mvk	Melavonate kinase
ENSMUSG00000041930	n/a
Trpv4	Transient receptor potential cation channel V4
NM_029992	n/a
Git2	G protein-coupled receptor kinase-interactor 2
NM_175120	n/a
Q9D1S6	n/a
NM_181075	n/a
Oasl2	2'-5'-oligoadenylate synthetase like protein 2
NM_028211	n/a
Tcf1	Hepatocyte nuclear factor 1-alpha
Usmg3	Upregulated during skeletal muscle growth 3
Rpl37	60S ribosomal protein L37
D5Ertd33e	n/a
Dnclc1	Dynein light chain 1, cytoplasmic
NM_029645	n/a
15E1_MOUSE	n/a
Cox6a1	Cytochrome c oxidase polypeptide VIa-liver
Pla2g1b	Phospholipase A2 precursor
NM_177759	n/a
NM_026886	n/a

Table 7.3: Genes within the *bv* candidate region for which full coding sequence has been obtained and no mutations have been identified. Those highlighted in **RED** were previously assessed as being good candidates for *bronx waltzer* (see Chapter 5).

Of the remaining 29 candidate genes, 24 are partially sequenced and 5 have no sequence available. These 5 consist of small, generally single-exon transcripts which were annotated subsequent to the second round of primer design and thus no attempt has been made to sequence them. As such, the next round of sequencing is likely to result in sequence coverage for these predicted transcripts. However, analysis of the evidence supporting their annotation (Table 7.4) suggests that most of them are unlikely candidates. Four out of the five are supported by only one cDNA clone, do not share similarities with ESTs from other species and do not match any protein domains in the protein families database Pfam (http://www.sanger.ac.uk/Software/Pfam/). This suggests that they may be the result of genomic contamination in the cDNA library, and therefore do not represent true transcripts. They are therefore very unlikely to be involved in the causation of the *bronx waltzer* phenotype. The exclusion of these four genes leaves 25 genes with partial sequence coverage which may harbour a mutation within their coding regions.

Predicted Gene	**Supporting evidence**	**Analysis**
Q8C864	1 mouse cDNA clone No hits in other species No Pfam domains predicted	Possible false transcript
ENSMUSG00000044804	9 mouse cDNA clones Hits in other species Pfam predicts ribosomal protein S12	Good supporting evidence, probably genuine transcript
Q8C4H9	1 mouse cDNA clone No hits in other species No Pfam domains predicted	Possible false transcript
Q9D2I7	1 mouse cDNA clone Weak hit to rat cDNA No Pfam domains predicted	Possible false transcript
Q9D4T7	1 mouse cDNA clones Weak hit to rat EST No Pfam domains predicted	Possible false transcript

Table 7.4: Summary of evidence in support of newly annotated, unsequenced transcripts. Data is collated from Ensembl entries, BLAST searches against the EST database and from Pfam.

7.4.3 Sequence polymorphisms

During the analysis of sequence data, a number of differences were identified between the *bronx waltzer* genetic background and the published mouse sequence derived from the strain C57Bl/6J, which either represent errors in the sequence reads or genuine polymorphisms between the two strains. These observations reinforce the importance of carrying out sequencing not only in the mutant *bv* mouse but also in the wild type congenic genome in order to eliminate the possibility of mistaking a polymorphism for a mutation. Future studies should utilise these data by carrying out sequencing of the putative polymorphism sites in 101/H, the strain used for the mapping backcross (see Chapter 3), in the hope that some may prove useful as markers in reducing the size of the critical region.

The second class of polymorphisms identified, those which differed between the *bv* mutant and wild type congenic genomes, represent an interesting finding. These 12 sites were extremely localised, with all of them occurring within sequence reads designed to amplify exons from the gene *Citron*. In every case, the *bv* allele matched that reported in the published mouse sequence derived from the strain C57Bl/6J (Waterston *et al.* 2002). Five of them were located within coding sequence and for each of these, the wild type allele matched the accessioned sequence for the gene (NM_007708), which was derived from the strain CD-1 (Holzman *et al.* 1994; Di Cunto *et al.* 2000).

Firstly, the identification of homozygous differences between the *bronx waltzer* mutant and congenic wild type DNA samples supports the assumption that the wild type sample is truly homozygous for the wild type allele, which is useful since the mouse from which the sample was taken did not quite reach the number of certifiable births required in order to be confidently pronounced +/+ (see Section 7.2.1).

More interesting is the possibility that the cluster of polymorphisms represents a localised region of non-recombination where the otherwise seemingly homogenous *bronx waltzer* genetic background has maintained some of the sequence variation introduced throughout successive generations of breeding. The *bv* mutation arose spontaneously on an unknown background in 1979 and has since been crossed to various other unrecorded strains. The colony used in this study has been maintained within a closed colony for 25 years, and although it is not "inbred" since brother-sister matings are not always possible, is likely to have become relatively homogenous, especially since several "bottle-necks" where few mice of breeding age were available have meant a reduction in genetic variation. This is supported by the evidence of this exon re-sequencing project, during which no other polymorphisms were identified between the two samples. Evidently, throughout the breeding process the original *bv* allele has been maintained and selected for over each generation, presenting the possibility that the localised region of polymorphisms identified around *Citron* may represent a portion of the *bronx waltzer* genome which has remained intact because it lies very close to the mutation and has thus been preserved in its original state, while the wild type alleles represent a strain to which the mutant mouse was crossed at some point in its history. In this scenario, the polymorphisms could be described as being in linkage disequilibrium with the mutation, that is, they co-segregate more frequently than might be expected if they were un-associated with it. In order to test this hypothesis, the simplest approach would be to carry out a screen of archived DNA samples from the *bronx waltzer* colony which have been collected over the years and establish whether the polymorphisms do indeed segregate with the mutation, effectively exploiting the background sequence variation to employ the breeding colony as a mapping cross.

It is also possible that the recombination events which gave rise to this localised region of sequence variation occurred before the mutation arose, and that the whole region has been conserved. However, the existence of 15 mice with recombinations within the candidate region defined by the *bv*/101 mapping cross, and the correlation of 1.77cM to 2.45Mb of physical distance (a ratio of 1:1.4, compared to the average 1:6 across mouse chromosome 5) suggests that this region generally undergoes recombination at relatively normal rates. This would imply that the localised region which has maintained these polymorphisms has done so independently of the rest of the candidate region, and its uniqueness is suggestive of an association with the *bv* locus.

It may seem beneficial to screen the inbred strain 101/H for these polymorphisms in the hope that the alleles may differ, making them useful as markers in the mapping cross to help narrow down the critical region. However, this would only be possible if it is found that the polymorphisms co-segregate with the mutation since any variability in the homozygous *bv*/*bv* mutant background would make it impossible to differentiate between that and the inbred strain. In this situation, where the polymorphisms prove to be co-segregating or even non-recombinant with the mutation throughout the breeding colony, the likelihood is that the markers would also be non-recombinant in the mapping cross and would thus add no new informative data. Nonetheless, in the absence of any other useful markers within the region, the nature of these should be fully investigated.

In order that this study can be more brought to completion more comprehensively, it is imperative that the quality of genome sequence and gene annotation within the region is improved. This would allow fully sequenced genes to be excluded more confidently since their full structure would be known, and may also bring to light novel genes which are as yet un-annotated and as such have not yet been sequenced. To this end, negotiations are underway to prioritise the finishing of the clones in the tile path already selected for HTGS, and to select and sequence new clones to fill the two tiling path gaps. Once this has been achieved, manual annotation of the region will be carried out and higher quality sequence coverage of the region will be made possible.