ExoSeq SNP Identification

Pre-processing

ExoTrace uses raw data, not ABI processed data, and applies minimal pre-processing prior to SNP detection. The pre-processing consists of a baseline correction, background removal and a mobility shift. This pre-processed raw trace is then base-called with PHRED (http://www.phrap.com/phred/). The process is illustrated in Figure 2, note the raw trace has not been mobility corrected so peaks in different channels are overlaid.

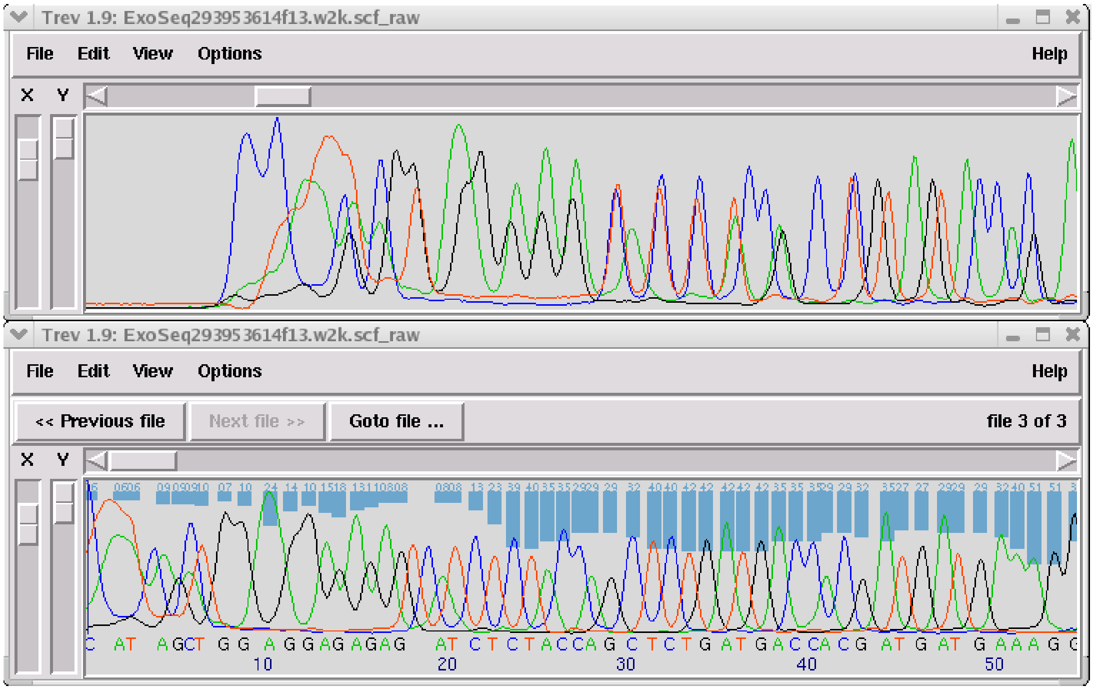

Trev 1.9: ExoSeq293953614f13.w2k.scf_raw
File
Edit
View
Options
Help
X
Y
Trev 1.9: ExoSeq293953614f13.w2k.scf_raw
File
Edit
View
Options
Help
<< Previous file
Next file >>
Goto file ...
file 3 of 3
X
Y
C AT AGCT G G A G G AG AGAG AT C T C T A C C A G C T C T G AT G A C CA C G AT G AT G A A A G C
10
20
30
40
50

Figure 2: Raw trace (top) and re-processed trace (bottom).

SNPs are only called over the part of the read that aligns to the reference sequence because peak heights are compared across a number of reads. Prior to SNP detection the reads are filtered and clipped to identify both bad reads and regions of the read where the trace quality it too low for SNP detection. Reads that do not align to the reference sequence and reads where the overall signal strength is too low are rejected. Once a read has been aligned to the reference sequence, the signal is extracted at the base positions for bases that align, thus producing a 'peak height' trace where each base corresponds to a base in the reference sequence. Figure 3 shows the peak height trace for the read shown in Figure 2.

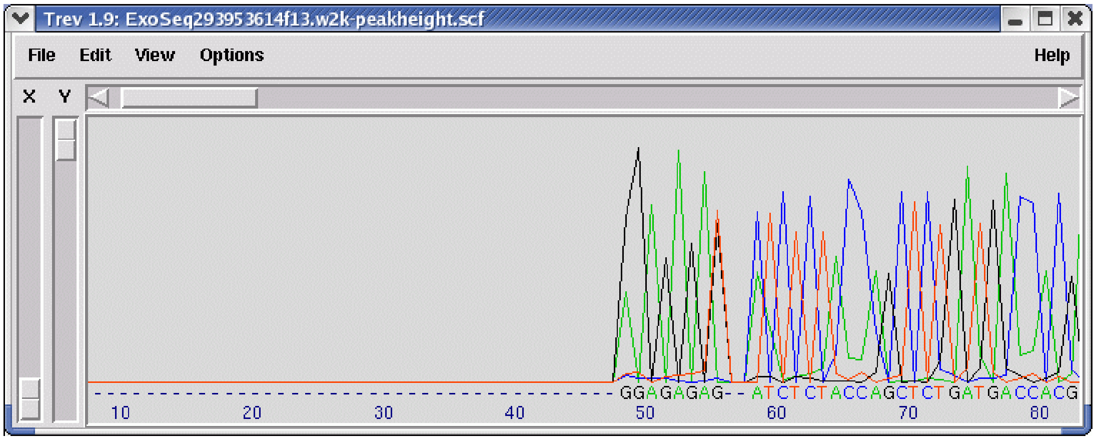
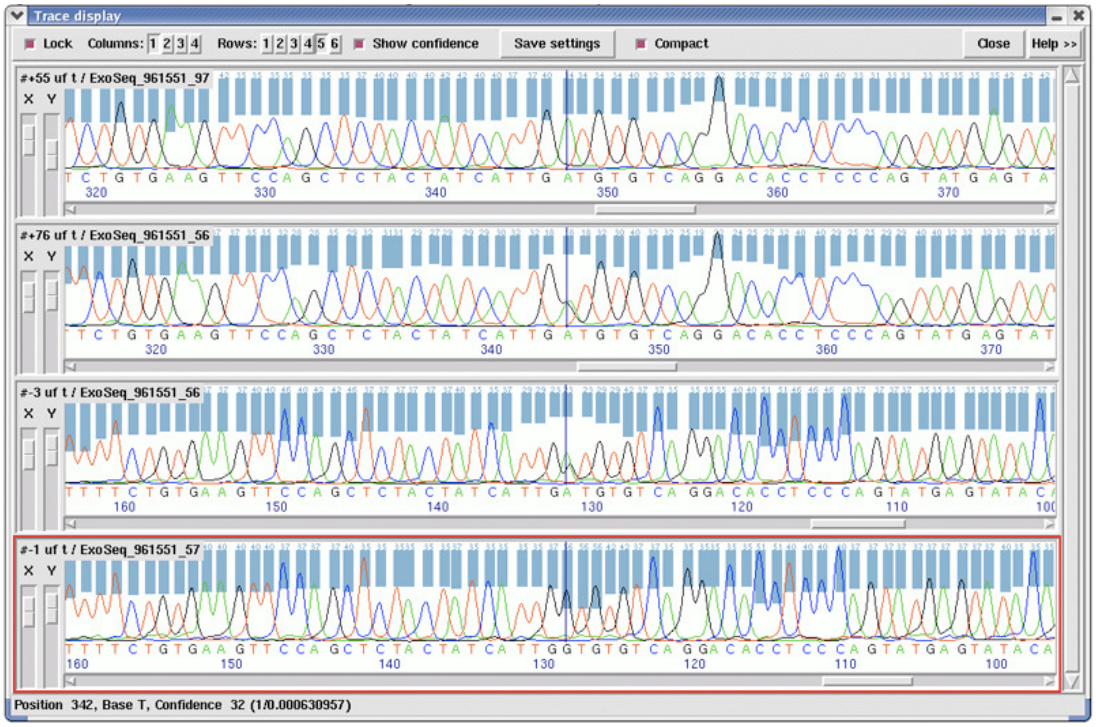

Trev 1.9: ExoSeq29395 3614f13.w2k-peakheight.scf
File
Edit
View
Options
Help
X
Y
GGAGAGAG- - ATCTCTACCAGCTCTGATGACCACG
10
20
30
40
50
60
70
80

Figure 3: Peak height trace corresponding to the read in Figure 2.

The number displayed at the bottom of the trace is an index into the reference sequence, for the read shown in Figure 3, the first base in the reference sequence that aligns to the read is base 48. For bases in the reference sequence that do not align to the read, the sequence is set to '-' and the peak height is set to 0. Since only peak height is used, the trace is made up of a series of sharp triangular shapes rather than the more familiar smooth peak shape. If the base is heterozygous two coincident peaks in separate channels will be seen. The two peaks can sometimes be slightly separated in the original trace due to the differing mobility between channels. In other cases there is only one clear peak and the second peak appears as a shoulder on a neighbouring peak, also the base position as identified by the basecaller may not correspond to a peak. In these cases the program looks for a peak or a shoulder in each channel close to the base position identified by the basecaller. An example of reads aligned to the reference sequence is shown in Figure 4.



Position 342, Base T, Confidence 32 (1/0.000630957)

Figure 4: Four reads from three different individuals aligned to the reference sequence.

Figure 4 shows four traces from three different individuals. The top two reads are the sense reads; note the base index, which is shown below each trace, increases to the right. The lower two reads are the antisense reads, which have been reverse-complemented hence the base index increases to the left. The middle two reads are from the same individual, notice the name displayed at the top of each panel (ExoSeq_961551_56) is the same, the trailing two digit number (56) being the anonymised individual identifier. The top and bottom reads are from two different individuals (ID 97 & 57). The shape of the sense trace is different to the shape of the antisense trace, for example look at the two G's to the right of the SNP. In both the antisense traces the peak heights are roughly equal whereas on both the forward traces the peak height of the second G is much larger. For this reason the sense and antisense reads are processed separately and later combined to produce the results. There is a SNP in the centre of the reads, identified by the vertical black line. These 3 individuals show all 3 possible genotypes, the top trace, a homozygous A and the bottom, a homozygous G. The middle two traces (from the same individual) show a combiniation of both and so are heterozygous.

SNP scoring

ExoTrace processes the sense and antisense sequence reads separately and subsequently combines the results to allow SNP scoring. An individual read is considered to be double stranded if the ExoTrace calls are the same for both the sense and antisense reads, whereas an individual is considered to be single stranded if ExoTrace can only call the base in one direction. The most common reason for only being able to call one read is because the other read failed or was of low quality. This is particularly true towards either end of the reference sequence where the read quality drops. To guarantee high quality data in both directions for the whole length of the exon under examination, an additional 75 base pair flank is added to the region of interest when designing primers (for more details please see the ExoSeq protocol on primer).

For each potential SNP identified by ExoTrace calls are gathered on all individuals and each SNP scored according to a pre-defined set of rules. Each SNP is assigned a status and placed in one of the following classes that have been ordered to reflect the confidence of the call.

Classes of SNPs called by the automatic SNP scoring program:

Class	Trusted	Description
SHW2	Y	All three possible genotypes observed - double stranded
MHo2	Y	Both possible homozygotes are seen double-stranded in multiple DNAs
MHe2	Y	One homozygote and the heterozygote are seen double-stranded in multiple DNAs
SHo2	Y	Both possible homozygotes are seen - one in only a single double-stranded DNA
She2	Y	One homozygote and the heterozygote are seen - one in only a single double-stranded DNA
SHW1	Y	All 3 possible genotypes seen - one only on single stranded DNA
ExoSeq threshold for automatic submission to dbSNP		
MHo1	N	Multiple examples of both possible homozygotes are seen but one is supported only by single-stranded DNA
MHe1	N	Multiple examples of one homozygote and the heterozygote are seen but one is supported only by stranded DNA
SHo1	N	Both possible homozygotes are seen - one in only a single single-stranded DNA
SHe1	N	One homozygote and the heterozygote - one in only a single single-

		stranded DNA
SCON	N	Single conflicting 'genotype'
MCON	N	Multiple conflicting 'genotypes'
REFO	N	Variation only seen between reference genome and reads

The most confident call is SHW2, for which there are one or more double stranded individuals for each of the three possible genotypes. The least confident is SHe1 for which a heterozygous SNP is seen on only one individual and in only one direction. A threshold score has been identified above which a SNP may be submitted to dbSNP without a manual check as indicated in Table 1. SNPs below this threshold need to be examined and manually confirmed prior to submission.

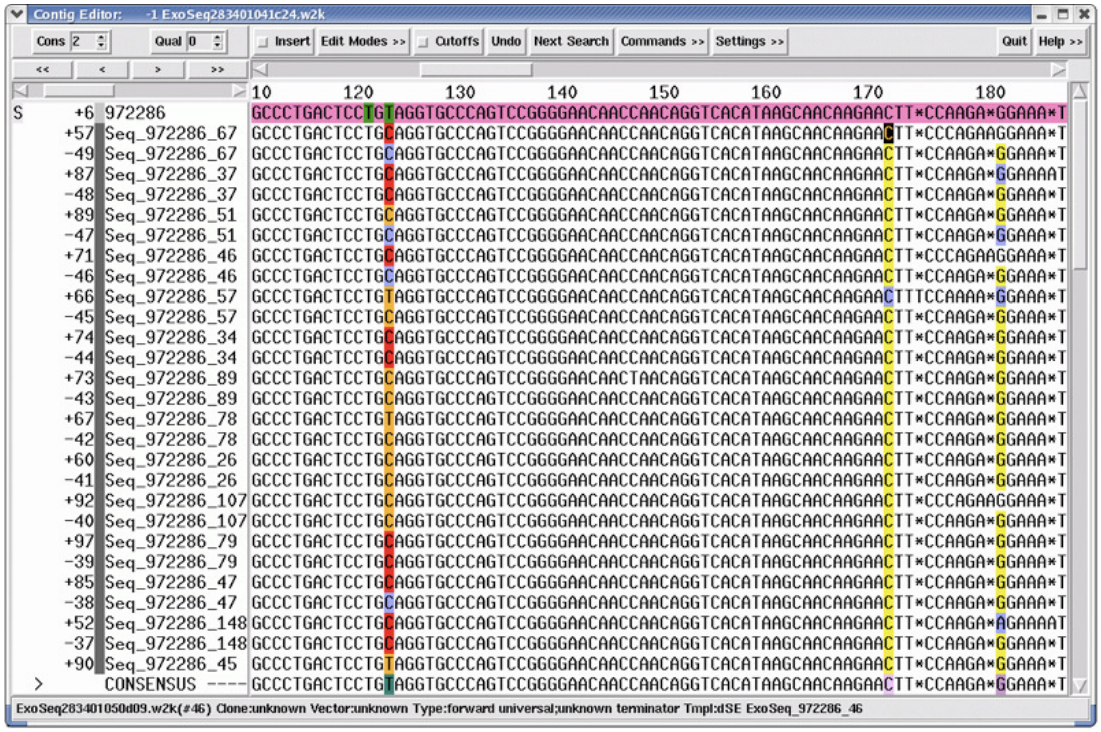
There are two statuses that are used to indicate a conflict, i.e. a base with conflicting calls for the same individual: CON1 where there are conflicting calls on a single individual, and CON2, where there are conflicting calls on multiple individuals.

The status REFO is used for SNPs that are called on the reference sequence only. This is a SNP where all individuals have the same homozygous base, either single stranded or double stranded, but this base differs from the reference sequence. Usually these occur when the reference sequence contains the rare allele or the reference sequence is incorrect. In some cases the reference sequence has been modified in a subsequent build and this SNP would not have been called if the most recent version of the genome was used.

A CHCK status indicates that a SNP is most likely to be real but should be checked manually. This could be a SNP for which there exists one or more conflicted individuals but which would pass the threshold for automatic submission if the reads with conflicting calls can be resolved manually. Most of these conflicts are caused by a single low quality read. In practice 75% of the SNPs scored as CHCK are confirmed manually, in contrast to SNPs scored as CON1 and CON2 for which a much smaller proportion can be confirmed manually.

Manual confirmation and review

The tool used for manual confirmation and review is a modified version of GAP4 (http://staden.sourceforge.net/), part of the Staden Sequence Analysis Package software, created for the ExoSeq project. The modified GAP4 database contains a single contig that matches the reference sequence, thus only reads that align to the reference sequence are included. SNPs called by ExoTrace may be viewed in the contig editor (Figure 5).



```
                         10        120       130       140       150       160       170        180
S          +6  972286           GCCCTGACTCCTGTAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +57 Seq_972286_67    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCCAGAAGGAAA*T
           -49 Seq_972286_67    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +87 Seq_972286_37    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAAAT
           -48 Seq_972286_37    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +89 Seq_972286_51    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -47 Seq_972286_51    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +71 Seq_972286_46    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCCAGAAGGAAA*T
           -46 Seq_972286_46    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +66 Seq_972286_57    GCCCTGACTCCTGTAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTTTCCAAAA*GGAAA*T
           -45 Seq_972286_57    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +74 Seq_972286_34    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -44 Seq_972286_34    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +73 Seq_972286_89    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACTAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -43 Seq_972286_89    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +67 Seq_972286_78    GCCCTGACTCCTGTAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -42 Seq_972286_78    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +60 Seq_972286_26    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -41 Seq_972286_26    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +92 Seq_972286_107   GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCCAGAAGGAAA*T
           -40 Seq_972286_107   GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +97 Seq_972286_79    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -39 Seq_972286_79    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +85 Seq_972286_47    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           -38 Seq_972286_47    GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +52 Seq_972286_148   GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*AGAAAAT
           -37 Seq_972286_148   GCCCTGACTCCTGCAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
           +90 Seq_972286_45    GCCCTGACTCCTGTAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
 >             CONSENSUS ----   GCCCTGACTCCTGTAGGTGCCCAGTCCGGGGAACAACCAACAGGTCACATAAGCAACAAGAACTT*CCAAGA*GGAAA*T
ExoSeq283401050d09.w2k(#46) Clone:unknown Vector:unknown Type:forward universal;unknown terminator Tmpl:dSE ExoSeq_972286_46
```

Figure 5: Modified GAP4 Contig Editor window.

At the top of the edit window is a copy of the reference sequence, here labelled with an internal database identifier (972286). The exon sequence is highlighted in pink on the reference sequence and previously known SNPs are tagged in dark green. The individual reads are aligned to and displayed below the reference sequence. The base numbering displayed is relative to the reference sequence. The reads are displayed in pairs for each individual, in this case all the read names have the form ExoSeq_972286_##, where ## is an internal database identifier for each individual. The sense reads are indicated by a '+' to the left of the read names and antisense reads by a '-'. The consensus sequence based on the individual reads is displayed at the bottom of the window.

A coloured tag is added to individual reads at those bases where ExoTrace has identified a SNP. The different colour tags seen on the reads are:

- Yellow 🟨 homozygous base which matches the reference sequence
- Orange 🟧 heterozygous SNP.
- Red 🟥 homozygous SNP.
- Blue 🟦 base that could not by called by ExoTrace.

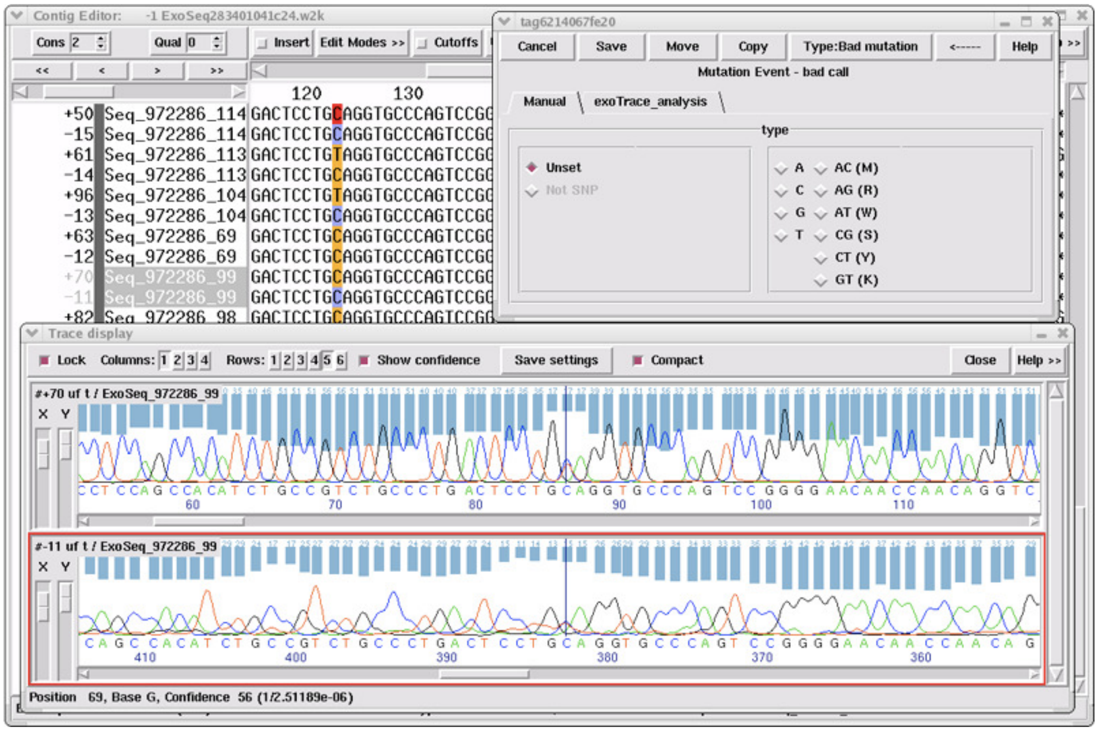
NB: ExoTrace may not be able to call a particular trace for a variety of reasons, for example, low signal strength or low PHRED quality scores. There are also a number of reads with no tag.
A different coloured tag is added to the consensus which gives the status for that SNP according to the SNP scoring program. The different colour tags seen on the consensus sequence are:

- Dark green 🟩 SHW2 All three genotypes seen on 2 strands.

Dark sea green	🟩	MHo2	Multiple Homozygotes seen on 2 strands.
Dark pink	🟪	SHo2	Single Homozygote seen on 2 strands.
Dark blue	🟦	MHe2	Multiple Heterozygote seen on 2 strands.
Dark red	🟥	SHe2	Single Heterozygote seen on 2 strands.
Bright green	🟩	SHW1	All three genotypes seen on 1 strand only.
Dark Lilac	🟪	Check.	
Bright turquoise	🟦	MHo1	Multiple Homozygotes seen on 1 strand only.
Bright blue	🟦	MHe1	Multiple Heterozygotes seen on 1 strand only.
Bright pink	🟪	SHo1	Single Homozygote seen on 1 strand only.
Bright red	🟥	SHe1	Single Heterozygote seen on 1 strand only.
Pale Lilac	🟪	Conflict on one template.	
Mid Lilac	🟪	Conflict on more than one template.	

These are a few reads that have no tag in Figure 5, these are either bad reads or in regions of the read which was excluded prior to SNP calling. These are usually regions where the quality is too low or the sequence differs significantly from the reference sequence. For instance looking to the left of the SNP at base180, on the sense read for individual 67 the sequence called by PHRED is CCCAGAA whereas the reference sequence is CCAAGA, i.e. the first A has been replaced by a C and these is an extra A called just to the left of the SNP. ExoTrace will not attempt to call a read at any base if the sequence differs from the reference sequence near that base.

It is possible to manually edit the SNPs called by ExoTrace or add new tags. This enables the conflicting calls for SNPs with a status of CHCK to be resolved. For instance at base 122 ExoTrace calls a heterozygous SNP (orange tag) on the sense read for individual 99 but is unable to call the antisense read (blue tag). If we display the traces (bottom left in Figure 6) the heterozygous SNP is clearly visible. The quality of the antisense read is too low in the region of the base for ExoTrace to call a SNP automatically but this is sufficient evidence for the tag to be edited manually. This is done using the GAP4 Tag edit box (top right in Figure 6).



```
                         120       130
+50 Seq_972286_114 GACTCCTGCAGGTGCCCAGTCCGG
-15 Seq_972286_114 GACTCCTGCAGGTGCCCAGTCCGG
+61 Seq_972286_113 GACTCCTGTAGGTGCCCAGTCCGG
-14 Seq_972286_113 GACTCCTGCAGGTGCCCAGTCCGG
+96 Seq_972286_104 GACTCCTGTAGGTGCCCAGTCCGG
-13 Seq_972286_104 GACTCCTGCAGGTGCCCAGTCCGG
+63 Seq_972286_69  GACTCCTGCAGGTGCCCAGTCCGG
-12 Seq_972286_69  GACTCCTGCAGGTGCCCAGTCCGG
+70 Seq_972286_99  GACTCCTGCAGGTGCCCAGTCCGG
-11 Seq_972286_99  GACTCCTGCAGGTGCCCAGTCCGG
+82 Seq_972286_98  GACTCCTGCAGGTGCCCAGTCCGG
```

Mutation Event - bad call

type

● Unset
○ Not SNP

○ A ○ AC (M)
○ C ○ AG (R)
○ G ○ AT (W)
○ T ○ CG (S)
○ CT (Y)
○ GT (K)

Position 69, Base G, Confidence 56 (1/2.51189e-06)

Figure 5: Manual editing in GAP4.

SNP Verification

1536 SNPs identified by ExoTrace including a range of SNPs from all different classes were chosen for verification with using a different experimental method, i.e. genotyping using the Illumina platform. Of these 1536 SNPs, 1365 produced results of which 102 were found to be not polymorphic on the Illumina platform. All of the 102 'non-polymorphic' SNPs were assayed on the Sequenom platform. Only 9/39 SNPs which gave results were found to be polymorphic, whilst 49 SNPs remain unconfirmed.

Differences in genotypes:

Genotypes	Number of SNPs	Percentage of SNPs
0	1119	82
1	111	8
>1	135	10