



Technology
Assessment
Report

DNA sequence data analysis

Automated mutation detection
using SoftGenetics®
Mutation Surveyor™ v2.51

September 2005

Authors:

Yogen Patel and Andrew Wallace

National Genetics Reference Laboratory (Manchester)
Regional Genetics Service
Central Manchester and Manchester Children's University Hospitals NHS Trust
St Mary's Hospital
Hathersage Road
Manchester M13 0JH
UK

www.ngrl.org.uk

Email:

yogen.patel@cmmc.nhs.uk
andrew.wallace@cmmc.nhs.uk

Funded by the United Kingdom:



Distributed to:



UK Genetic Testing Network

Distributed to EMQN registered members:



- *The authors declare that they have no competing financial interests.*
- *SoftGenetics®, the software developer of this product, has been given the right to comment on this assessment.*
- *This report has been peer reviewed.*
- *The UK Department of Health does not necessarily endorse or accept the conclusions of this report.*
- *© September 2005 National Genetics Reference Laboratory (Manchester).*

Table of Contents

1.	TITLE	3
2.	ABSTRACT	3
3.	INTRODUCTION	4
3.1.	Purpose of the Study	4
3.2.	Overview of the product	4
3.3.	Summary of Manufacturer's claims	5
3.4.	Additional features	5
3.5.	User Interface	5
3.6.	Main Output files	10
3.7.	Analysis and Output measures	12
3.7.1.	Parameters used by Mutation Surveyor for mutation scanning	12
3.7.2.	Additional Mutation Surveyor outputs measures (tabulated output)	12
3.8.	Analysis procedure adopted for this study	13
3.9.	Default analysis settings	14
3.10.	Illogical data	14
4.	PERFORMANCE TESTING OF MUTATION SURVEYOR	16
4.1.	EMQN, DNA sequencing EQA scheme data	17
4.1.1.	Data Source	17
4.1.2.1.	Participants	17
4.1.2.2.	Structure of the scheme	17
4.1.2.3.	Instrumentation and sequencing chemistry	17
4.1.3.	Analysis settings	18
4.1.4.	Timing and work load	18
4.1.5.	Mutation Detection	20
4.1.5.1.	False negatives broken down by directional coverage	20
4.1.5.2.	Comments	20
4.1.5.3.	Screen snapshots of uni-directional false negatives	21
4.1.5.4.	Screen snapshot of uni-directional false positive	23
4.2.	UK CMGS, comparative study of diagnostic sequencing data	23
4.2.1.	Data Source	23
4.2.2.1.	Participants	23
4.2.2.2.	Structure of the scheme	23
4.2.2.3.	Instrumentation and sequencing chemistry	24
4.2.3.	Analysis settings	26
4.2.4.	Timing and work load	26
4.2.5.	Mutation Detection	26
4.2.5.1.	False negatives broken down by directional coverage	26
4.2.5.2.	Comments	27
4.2.5.3.	Screen snapshot of true bi-directional false negatives	28
4.2.5.4.	Screen snapshots of true uni-directional false negatives	28
4.3.	VariantSEQr™ resequencing data set	30
4.3.1.	Data Source	30
4.3.1.1.	Composition and coverage of the NF2 VariantSEQr™ kit data	30
4.3.1.2.	Instrumentation and sequencing chemistry	30
4.3.3.	Analysis settings	30
4.3.4.	Timing and work load	31
4.3.5.	Mutation Detection	31
4.3.5.1.	False negatives broken down by directional coverage	31
4.3.5.2.	Comments	31
4.3.5.3.	Screen snapshots of true bi-directional false negatives	33
4.3.5.4.	Screen snapshots of true uni-directional false negatives	34
4.4.	NF2 Exon linked sequencing data set	35
4.4.1.	Data Source	35
4.4.1.1.	Composition and coverage of the NF2 exon linked data	35
4.4.1.2.	Instrumentation and sequencing chemistry	35
4.4.3.	Analysis settings	35
4.4.4.	Timing and work load	36
4.4.5.	Mutation Detection	36
4.4.5.1.	False negatives broken down by directional coverage	37
4.4.5.1.1.	Bi-directional false negatives	37
4.4.5.1.2.	Uni-directional false negatives	39
5.	OTHER FEATURES OF MUTATION SURVEYOR	44
5.1.	Contig alignment, reference sequences and automated mutation naming in Mutation Surveyor	44
5.1.1.	NF2 VariantSEQr data set	44
5.1.2.	NF2 Exon linked data set	46
5.2.	Strengths of Mutation Surveyor	47
5.3.	Limitations of Mutation Surveyor	47
6.	SUPPLIER DETAILS AND CURRENT PRICES	51
7.	SUMMARY	52
8.	ABBREVIATIONS / GLOSSARY	54
9.	APPENDIX	54

An evaluation of automated mutation detection using SoftGenetics® sequence data analysis software Mutation Surveyor™ v2.51

Yogen Patel and Andrew Wallace

2. ABSTRACT

As DNA sequencing is increasingly important within diagnostic laboratories and manual methods of data analysis are labour intensive, diagnostic laboratories have a need for accurate and rapid automated mutation detection that can perform in a clinical diagnostic setting.

The aim of this study was to assess the performance of automated mutation detection using SoftGenetics® sequence data analysis software Mutation Surveyor™ v2.51 in a diagnostic setting. We tested four sets of bi-directional sequence data that covered a broad spectrum of sequencing chemistries, laboratories, sequencing platforms and read lengths, attempting to represent the range of bi-directional sequence data generated in clinical diagnostic laboratories.

In bi-directional mode, Mutation Surveyor is claimed to detect >99% of mutations, with sensitivity to the mutant allele extending down to 5% of the primary peak provided sequence quality meets a minimum Phred score of 20¹. Since Mutation Surveyor does not provide Phred quality scores it is unclear whether all of the data we used meets this requirement. However, in the four data sets in this study after excluding all possible explanations for false negative results through visual inspection of the trace data, the bi-directional false negative rate ranged from 0.0-4.9% depending on data set.

Although Mutation Surveyor claims reliable detection of 5% mosaic mutations we found that sensitivity was depressed even further and only 62% (33/53) of mosaic mutations were detected under default automated mutation detection settings.

Mutation Surveyor showed decreased sensitivity and an increased false positive rate on data produced by the Beckman CEQ8000 platform using the CEQ DTCS chemistry

Mutation Surveyor was able to de-convolute 89% (155/175) of heterozygote indel mutations into separate alleles. However separation into the two alleles does not permit the automated detection of mutations downstream of the indel as Mutation Surveyor failed to do this on all 49 cases where a mutation lay downstream of an indel in our data set.

Although Mutation Surveyor classifies sequence data quality, the mutation detection rate is depressed in data classified as acceptable when the program is run under the default settings.

Mutation Surveyor was able to correctly name the majority of mutations detected according to the reference sequence used. Mutation Surveyor did have difficulty in naming frameshift mutations sequenced in the reverse orientation, the names designated were often displaced upstream of the accepted name for the mutation.

Although Mutation Surveyor has a facility for designating a region of interest (ROI) on the reference sequence and the mutation output table indicates the start and end points of analysis, it is left to the user to reconcile the two measures to determine whether the ROI has been adequately covered for a given sample.

Nevertheless, Mutation Surveyor is a very comprehensive and useful program for detection of mutations in DNA sequence data and can make a very significant contribution in diagnostic laboratories in helping to ease the burden of sequence data analysis. Although we have highlighted weaknesses with the program when it is used in automatic mode with default settings 'out of the box', the user has the facility to alter many parameters which could increase overall sensitivity by tailoring the mutation detection algorithm to local sequencing chemistry, strategy etc.

Clearly it is beyond the scope of this Technology Assessment to test all possible configurations. The manufacturer could address these problems by recommending possible configurations of Mutation Surveyor to suit diagnostic laboratories which further minimised the likelihood of false negatives.

¹ In communication with SoftGenetics they say this is the minimum requirement to achieve their claimed detection rates, although this is not made clear in their Mutation Surveyor operation manual.

3. INTRODUCTION

3.1. Purpose of the Study

As sequencing technologies improve and the cost of sequencing reagents falls over time, direct sequencing as a primary mutation scanning technique is becoming more viable. However, diagnostic laboratories using direct sequencing for mutation scanning are hindered by the large amounts of sequence data that are generated and which needs careful analysis. Analysis has become the bottleneck in sequence based mutation scanning in diagnostic laboratories as it needs to be carried out carefully and the data is typically checked by two separate individuals. Consequently there is a need for software and systems that improve efficiency of DNA sequence analysis whilst capitalising on the technique's sensitivity.

The requirements of diagnostic laboratories for DNA sequence analysis programs that are applied to mutation scanning are however particularly stringent. False positives can be tolerated provided they are not excessively frequent as these can be differentiated by manual review. False negatives are however not tolerated by diagnostic users. If a diagnostic user cannot be certain that mutations have been excluded in DNA sequence data scanned automatically then the only option is to undertake a manual review of the data, thus negating any time savings. The presence of even a small proportion of false negatives consequently renders the value added by a mutation detection algorithm to almost zero. The diagnostic user is also looking for a whole suite of other features from any sequencing analysis software which although not individually as important as the core mutation detection algorithm, are collectively so.

There are a number of commercial sequence analysis software packages that may fulfil these functions, which could be applied to DNA sequence analysis in diagnostic laboratories. Given that sequencing is a routine task within diagnostic laboratories and manual methods of data analysis are labour intensive, diagnostic laboratories are awaiting with interest an assessment of the performance of mutation detection applications such as Mutation Surveyor™ in a clinical diagnostic setting.

The aim of this study is to assess the performance of automated mutation detection using SoftGenetics® sequence data analysis software Mutation Surveyor™ v2.51 in a diagnostic setting. We tested four sets of sequence data that covered a broad spectrum of sequencing chemistries, laboratories, sequencing platforms and read lengths. Two of the data sets were drawn from data generated in a number of different diagnostic laboratories using different sequencing chemistries and platforms, whereas the other two were generated internally. We have attempted to represent the spectrum of sequence data generated in clinical diagnostic laboratories.

3.2. Overview of the product

Mutation Surveyor™ performs mutational analysis on both bi-directional and uni-directional DNA sequence data and can be used in both mutation scanning and genotyping studies. DNA sequence data is analysed for potential variants using an anti-correlation method which performs a comparison of the test data to control data. Homozygote and heterozygote mutations are represented graphically as anomalies in a mutation electropherogram and are automatically flagged once a threshold is exceeded.

In bi-directional mode, Mutation Surveyor is claimed to detect >99% of mutations, with sensitivity to the mutant allele extending down to 5% of the primary peak. In uni-directional mode a detection rate of $\geq 95\%$ with a mutant allele sensitivity of 10% is quoted.

In addition to being sensitive to the presence of insertions and deletions Mutation Surveyor also de-convolutes heterozygote insertion deletion mutations (indels) of 1-100 bp in size into both alleles thus permitting continuation of the mutational analysis downstream of the indel. Mutation Surveyor also automatically aligns, assembles contigs, and performs mutational analysis on up to 400 lanes of data simultaneously (a 48 lane capacity version of the software is also available), is capable of fully unattended operation and can process approximately 1 billion base pair (bp) of sequence data in a day

3.3. Summary of Manufacturer's claims

[see <http://www.softgenetics.com/ms/index.htm> for further details]

- The program performs equally well with either terminator or primer chemistries from either gel or capillary systems from all manufacturers of DNA sequencing instrumentation
- Compatible with .scf, .ab1 and .abi sequence data file formats
- Capable of aligning sequences regardless of sequence quality or text call accuracy
- Enhanced indel detection - identifies heterozygous insertion/deletion mutations which often require manual review by trained personnel
- Detection of mutations downstream of a de-convoluted frameshift
- Mutation detection sensitivity down to 5% of the primary peak
- Detection of both the homozygous/heterozygous allele states
- Automated and accurate naming of mutations called - mutation names are compliant with the nomenclature recommendations produced by the Human Genome Variation Society (HGVS; www.hgvs.org).

3.4. Additional features

- GenBank/Sequence file editor
- File renaming feature – permits bulk file renaming
- Forward and reverse file pairing and batching
- An autorun feature to automatically run files within pre-defined folders
- Log file generator and editor to define file locations for autorun batch processing
- Sequence data file format conversion (from .ab1/.abi/ .fasta files to .scf files)
- Create .scf synthetic sequence data files (from .seq/.gbk/.scf/.ab1/.abi files)
- Ability to set user defined ROIs on the GenBank reference sequence and configure output measures to only present data from these regions or from the CDS regions (plus portions of intronic sequence either side) as given in the GenBank file.

3.5. User Interface

To illustrate the software's interface the screenshots below show the program's main graphical interfaces (these sequentially follow the processing of a set of 5 patient samples - for a single exon fragment) :

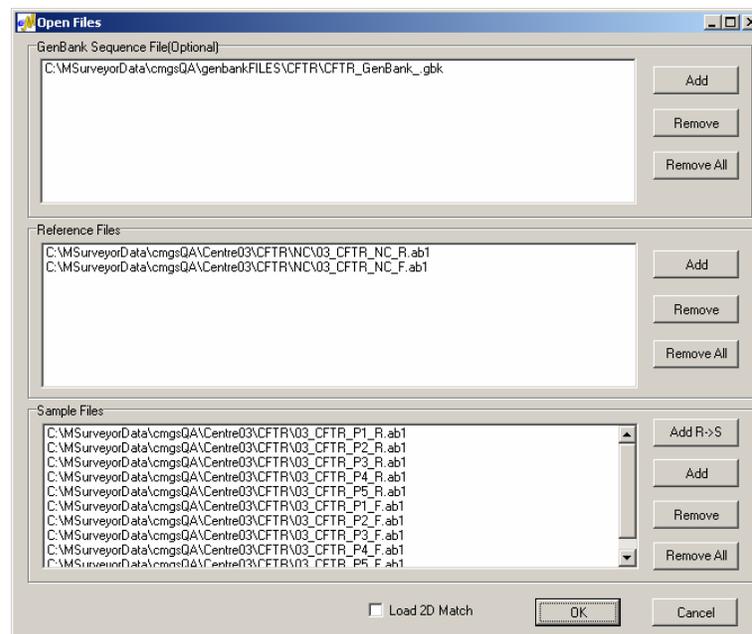


Figure 1. The main data-input window [prompts the user to specify the location of the three different input files; a GenBank reference file, wild type (WT) normal control traces and the test sample trace files]

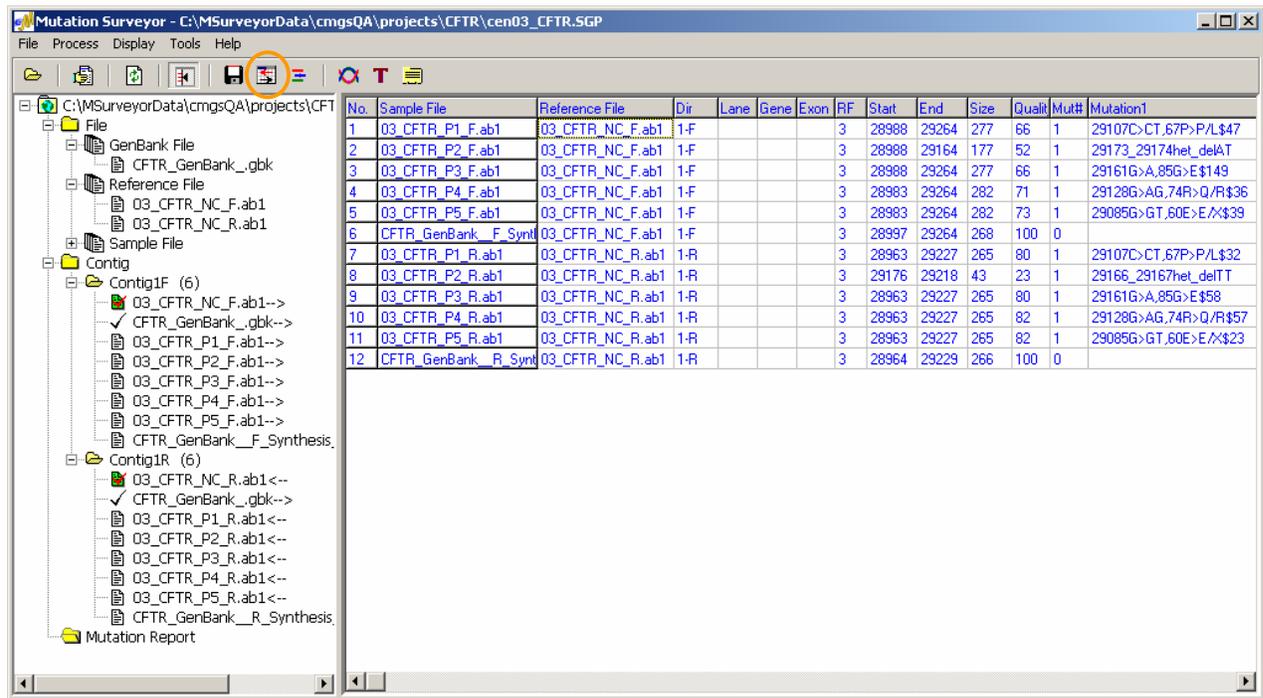


Figure 2. Main project window presented by Mutation Surveyor once data analysis is complete [the left-hand pane shows all the input files in a directory/folder structure . The samples are subdivided into contigs and the appropriate GenBank file and WT normal controls are highlighted with a tick; the blue text within the main results pane are hyperlinks which when clicked show the analysed sequence data in a graphical form; circled in orange is the bi-directional output button - see figure 3]

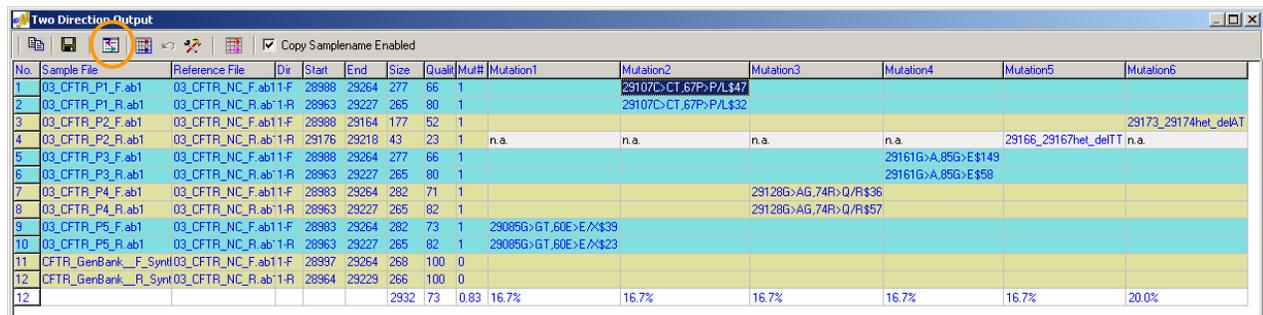


Figure 3. Bi-directional output table [presents the results from the main project window shown in figure 2 as paired forward and reverse reads; circled in orange is the advanced bi-directional output button - see figure 4]

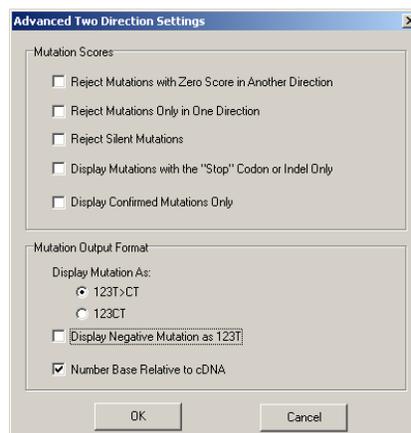


Figure 4. Settings for the advanced bi-directional output [filtering and nomenclature output options for the data from the bi-directional output window shown in figure 3 above]

No.	Sample File	Reference File	Dir	Start	End	Size	Qual#	Mut#	Mutation1	Mutation2	Mutation3	Mutation4	Mutation5	Mutation6
1	03_CFTR_P1_F.ab1	03_CFTR_NC_F.ab1	1-F	28988	29264	277	66	1		c.200C>CT.p.P67PL\$47 c.200C>CT.p.P67PL\$37				
2	03_CFTR_P1_R.ab1	03_CFTR_NC_R.ab1	1-R	28963	29227	265	80	1						
3	03_CFTR_P2_F.ab1	03_CFTR_NC_F.ab1	1-F	28988	29164	177	52	1						c.266_267het_delAT
4	03_CFTR_P2_R.ab1	03_CFTR_NC_R.ab1	1-R	29176	29218	43	23	1	n.a.	n.a.	n.a.	n.a.	c.259_260het_delTT	n.a.
5	03_CFTR_P3_F.ab1	03_CFTR_NC_F.ab1	1-F	28988	29264	277	66	1				c.254G>A.p.G85E\$149		
6	03_CFTR_P3_R.ab1	03_CFTR_NC_R.ab1	1-R	28963	29227	265	80	1				c.254G>A.p.G85E\$58		
7	03_CFTR_P4_F.ab1	03_CFTR_NC_F.ab1	1-F	28983	29264	282	71	1		c.221G>AG.p.R740R\$36				
8	03_CFTR_P4_R.ab1	03_CFTR_NC_R.ab1	1-R	28963	29227	265	82	1		c.221G>AG.p.R740R\$57				
9	03_CFTR_P5_F.ab1	03_CFTR_NC_F.ab1	1-F	28983	29264	282	73	1	c.178G>GT.p.E60E\$39					
10	03_CFTR_P5_R.ab1	03_CFTR_NC_R.ab1	1-R	28963	29227	265	82	1	c.178G>GT.p.E60E\$23					
11	CFTR_GenBank_F_S03_CFTR_NC_F.ab1		1-F	28997	29264	268	100	0						
12	CFTR_GenBank_R_S03_CFTR_NC_R.ab1		1-R	28964	29229	266	100	0						
12				2932	73	0.83	16.7%		16.7%	16.7%	16.7%	16.7%	16.7%	20.0%

Figure 5. Advanced bi-directional output table [presents the results in the bi-directional output table (see figure 3 above) with the selected filtering option and cDNA numbering according to the GenBank file; note how the mutations are named by the software at the cDNA level followed by the protein interpretation. A \$ symbol then precedes the mutation score generated by the software; the mutation circled in orange is illustrated in the hyperlinked graphical output below - see figure 6 below]

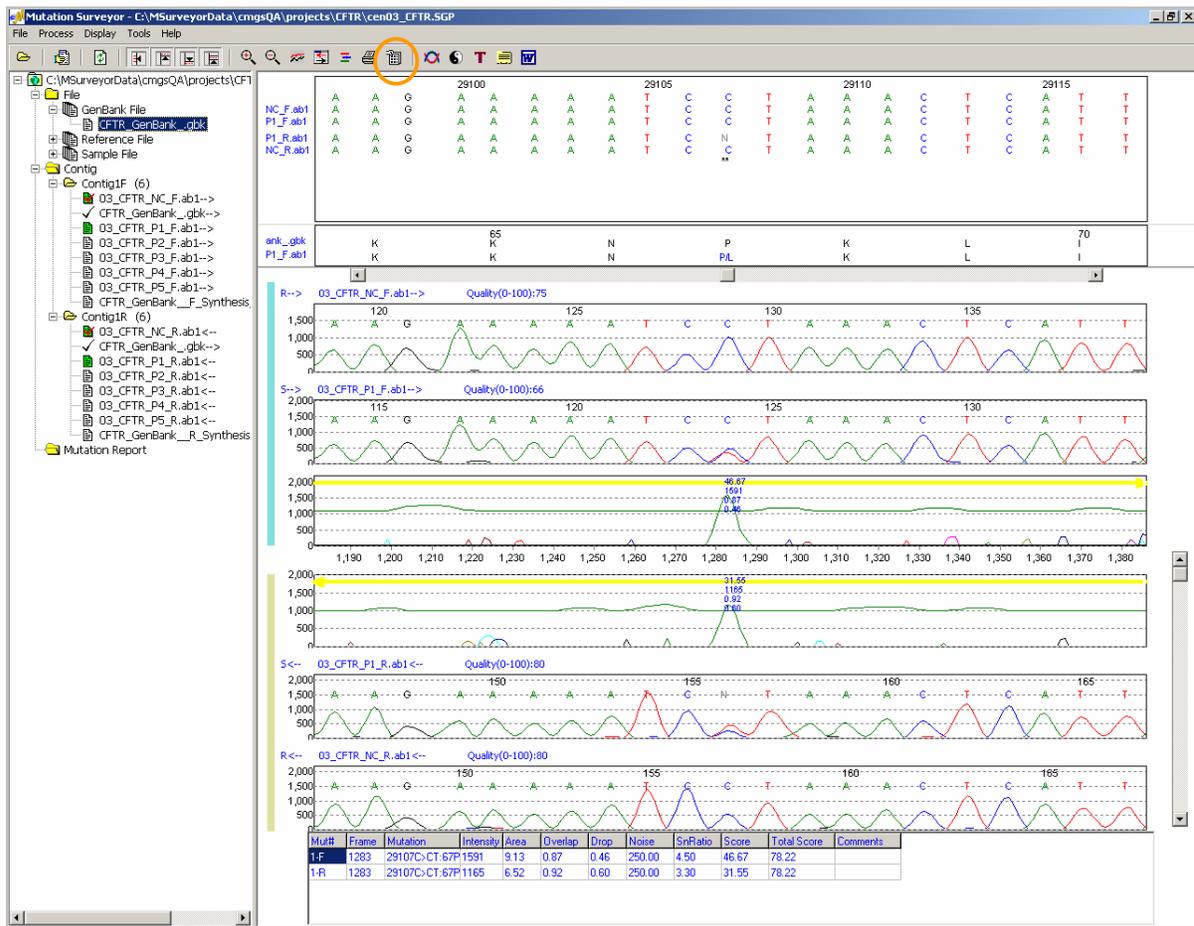


Figure 6. Graphical output of analysed sequence data part 1 [the upper right-hand pane shows the GenBank sequence as text, normal control (NC) sequence and the test sample sequence, below the nucleotide sequence is the amino acid sequence; in the lower right-hand pane forward traces are highlighted by a blue bar; reverse traces by a cream bar; the middle electropherogram is the mutation electropherogram which highlights any differences between normal control and reference traces as spikes; circled in orange is the print clinical report button - see section 3.6 Main output files]

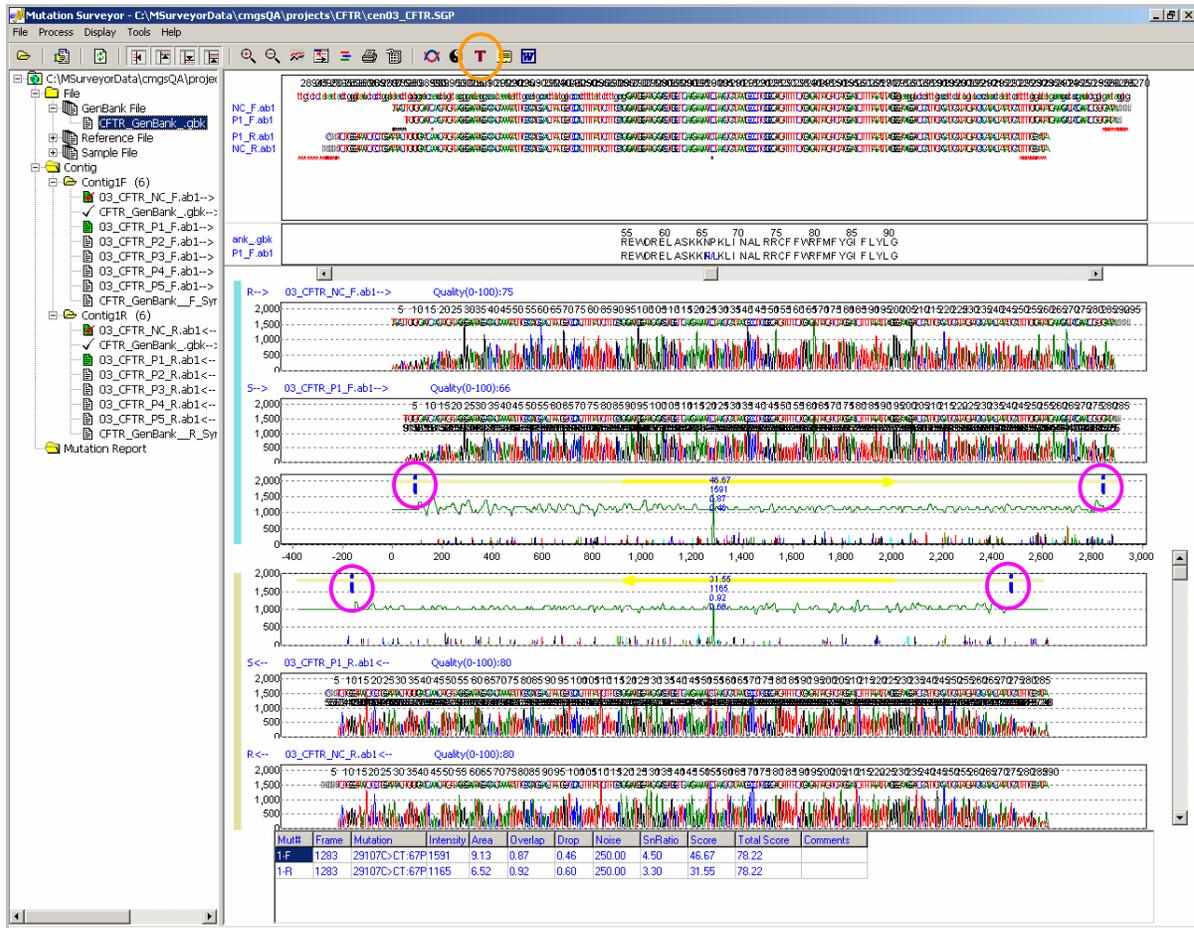


Figure 7. Graphical output of analysed sequence data part 2 [same data as in figure 6 above but zoomed out to cover the whole sequence electropherogram; circled in orange is the text output button - see figure 8 below; circled in pink are small vertical blue bars, these represent the start and end points of the quality read length/size as given in the tabulated output see figures 3/5 and section 3.7.2. *Additional Mutation Surveyor outputs*]

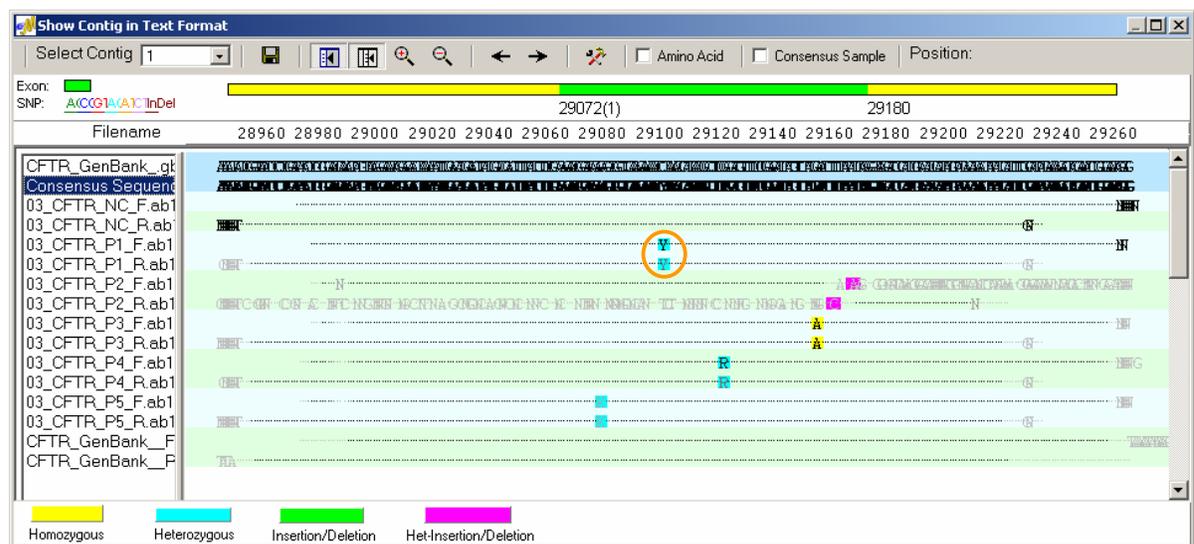


Figure 8. Sequence text output [all the fragments in this test are aligned to the GenBank sequence and the identified mutations highlighted accordingly; circled in orange a C>T substitution - see figure 9 below]

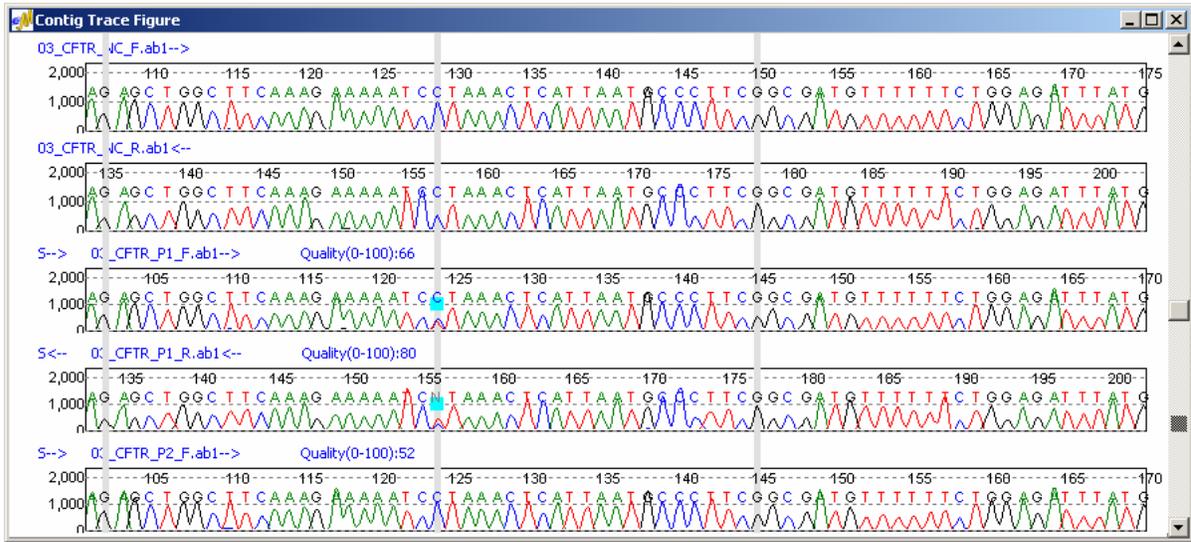


Figure 9. Trace data illustrating a mutation – hyperlinked from the sequence text output window (see figure 8)

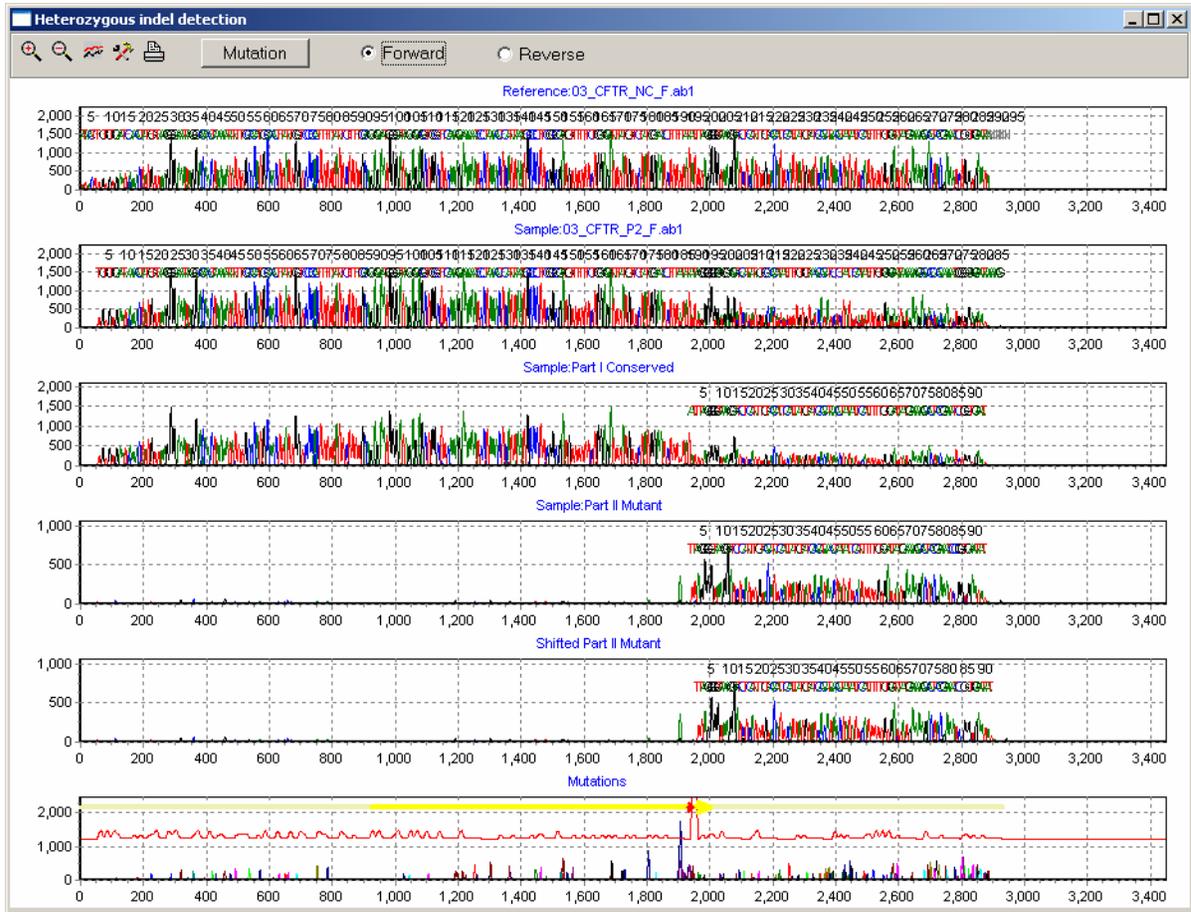


Figure 10. Sample graphical output of a heterozygous TT deletion part 1 [the whole fragment is displayed in the forward orientation with de-convolution after the heterozygous deletion]



Figure 11. Sample graphical output of a heterozygous TT deletion part 2 [same data as shown in figure 10 with the view zoomed in to show greater detail]

3.6. Main Output files

The results of mutation analysis are reported in a concise tabular format (figure 2). The reports within the output project file are hyperlinked to sequence electropherograms and in this window variants are highlighted by peaks rising above a threshold in the mutation electropherogram (figures 6 and 7).

The results reported in any of the tabulated formats can either be copied and pasted in to other applications or exported in tab delimited text, Excel, XML or HTML formats. However none of the exported formats maintain hyperlinks to graphical outputs and thus can only be viewed as standalone tables. To fully view data presented by a Mutation Surveyor analysis the user must run the saved project file in a functional version of the Mutation Surveyor software.

The user is also able to print out a Clinical Report, from within the graphical view (figure 12). A clinical report is formatted such that the page header contains the sample information. The report can also be set to print all listed mutations in the GenBank file and/or individual mutations detected in the analysis. In addition the header information can be personalised to suit the user or laboratory, however this feature is limited e.g. a full laboratory letterhead can not be added.

When several DNA fragments generated by different primer pairs are being analysed simultaneously for a patient specimen the 'Whole Gene Output' format is useful. Here bi-directional data for multiple fragments are grouped by specimen name (figure 12.1). In 'Whole Gene analysis', the sequences covered by the different primer pairs may overlap. Overlapping regions can serve as internal controls - if a mutation is real, it should be present in a specific region, regardless of primer set. Therefore, if a mutation is found at point X with one primer set yet is not found with a second primer set, the mutation may be a false positive. Consequently, when overlaps occur, mutation detection in the overlapping region is claimed to be very accurate. The Whole Gene Output Table is accessed by its icon, and is available only in the bi-directional output window.

There are a number of alternative views available within Mutation Surveyor to display results and summarise data graphically that have not been illustrated in this report. Those illustrated were judged to be those most suited to the types of data analysed in the preparation of this report.

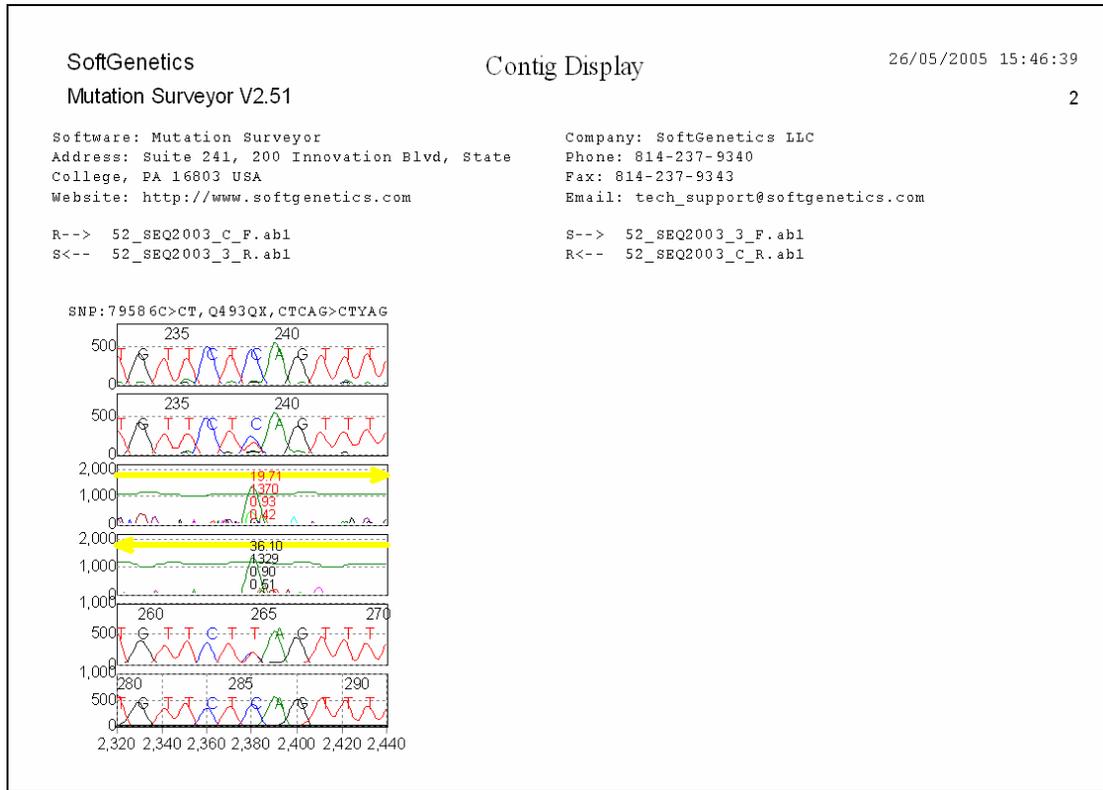


Figure 12. A sample clinical report.

The screenshot shows a 'Whole Gene Output' table with the following columns: No., Sample File, Reference File, Dir, Exon, RF, Start, End, Size, Quality, Mut#, Mutation1, and Mutation2. The table contains 23 rows of data, each representing a different sample and its corresponding reference file, exon, and various metrics.

No.	Sample File	Reference File	Dir	Exon	RF	Start	End	Size	Quality	Mut#	Mutation1	Mutation2
1	S19_frg01_F2_G04	S36_frg1_F_D05_023-F		1	-39	509	548	28	0	n.a.	n.a.	
2	S19_frg01_R2_G04	S36_frg1_F_D05_023-R		1	-88	480	568	31	0	n.a.	n.a.	
3	S19_frg02_F2_A05	S36_frg1_F_D05_023-F		1	101	478	378	11	0	n.a.	n.a.	
4	S19_frg02_R2_A05	S36_frg1_R_D05_023-R		1	23	485	463	18	0	n.a.	n.a.	
5	S19_frg03_F2_D02	S36_frg03_F2_H08_7-F		1	405	853	449	19	0	n.a.	n.a.	
6	S19_frg10_F_C03_0	S36_frg10_F_D05_015-F		1	2649	3179	531	26	0	n.a.	n.a.	
7	S19_frg10_R_C03_0	S36_frg10_R_D05_015-R		1	2623	3152	530	38	0	n.a.	n.a.	
8	S19_frg11_F_C03_0	S36_frg11_F_D05_017-F		1	2939	3276	338	18	0	n.a.	n.a.	
9	S19_frg12_F_C03_0	S36_frg12_F_D05_08-F		3	3375	3756	382	30	0	n.a.	n.a.	
10	S19_frg12_R_C03_0	S36_frg12_R_D05_08-R		3	3334	3734	401	43	0	n.a.	n.a.	
11	S19_frg13_F_C03_0	S08_frg13_F_H01_016-F		1	3668	4111	444	27	0	n.a.	n.a.	
12	S19_frg13_R_C03_0	S08_frg13_R_H01_016-R		1	3637	4073	437	36	0	n.a.	n.a.	
13	S19_frg14_F_C03_0	S36_frg14_F_D05_021-F		1	4024	4331	308	19	0	n.a.	n.a.	
14	S19_frg14_R_C03_0	S36_frg14_R_D05_021-R		1	3997	4300	304	33	0	n.a.	n.a.	
15	S19_frg15_F_C03_0	S08_frg15_F_H01_011-F		1	4294	4699	406	28	0	n.a.	n.a.	
16	S19_frg15_R_C03_0	S08_frg15_R_H01_011-R		1	4269	4667	399	44	0	n.a.	n.a.	
17	S19_frg16_F_C03_0	S08_frg16_F_H01_013-F		1	4700	5126	427	23	0	n.a.	n.a.	
18	S19_frg16_R_C03_0	S08_frg16_R_H01_013-R		1	4628	5093	466	34	0	n.a.	n.a.	
19	S19_frg17_F_C03_0	S36_frg17_F_D05_023-F		1	5089	5492	404	18	6	n.a.	n.a.	
20	S19_frg17_R_C03_0	S36_frg17_R_D05_023-R		1	4991	5403	413	29	2	n.a.	n.a.	
21	S19_frg18_F_C03_0	S36_frg18_F_D05_03-F		1	5331	5860	530	28	0	n.a.	n.a.	
22	S19_frg18_R_C03_0	S36_frg18_R_D05_03-R		1	5303	5841	539	43	0	n.a.	n.a.	
23	S19_frg19_F_C03_0	S08_frg19_F_H01_014-F		1	5804	6173	370	24	0	n.a.	n.a.	

Figure 12.1. A sample 'Whole Gene Output' table for a single specimen.

3.7. Analysis and Output measures

By performing a comparison of the actual raw sequence traces, SoftGenetics' Mutation Surveyor is claimed to offer significantly enhanced mutation detection sensitivity over sequence text comparison programs, which are prone to high levels of false positives and vary in sensitivity dependent upon the quality of the original sequence and the accuracy of the base call program.

Mutations are called on a comparison of a test sample trace to a reference control trace and is dependent on the satisfaction of the following parameters; mutation height, mutation score, signal to noise ratio, overlapping and dropping (see section 3.7.1 for a full description of these parameters).

3.7.1. Parameters used by Mutation Surveyor for Mutation Scanning :

- *Mutation peak height*; is the maximum height of the mutation peak in the mutation electropherogram.
- *Signal to noise ratio*; where noise is the median peak height of all the minor mutation peaks (in the mutation electropherogram) within a local region. The signal to noise ratio is used to determine the confidence of the peaks, where the confidence is calculated using a Gaussian distribution, assuming that the median value (σ) is the noise and the highest value is the signal. The area of the Gaussian curve under 1σ is 68%, 2σ is 95%, and 3σ is 99.7%. The error probability of the mutation peak is 1 - confidence.
- *Overlapping factor*; is a measure of relative shift of the two peaks at the mutation position in the horizontal direction. The overlapping factor calculates the horizontal, overlapping percentage of a wild type peak to the mutant peak.
- *Dropping factor*; the drop in height of the normal peak at the position of the mutation relative to the neighbouring peaks.
- *Mutation score*; is derived from the signal to noise ratio, the dropping factor and overlapping factor expressed as :

$$= -10 \log (\text{error probability}) = -10 \log \left[\text{erfc} \left(\frac{s/n}{\sqrt{2}} \right) * \text{dropping factor} * \text{overlapping factor} \right]$$

(where *signal* is defined as a peak in the mutation electropherogram and *noise* is defined as the smaller peaks surrounding the mutation in the electropherogram; *erfc* (x) is the complementary error function)

Accuracy is defined as 100% minus the error percentage, where the highest possible confidence, 99.9% corresponds to a mutation score of 30. A score of 20 corresponds to 99% accuracy; a score of 10 corresponds to 90% accuracy.

- *Quality Trim*; is a default overrideable option in the Mutation Surveyor settings. Quality trim cuts the poor quality sequence from the beginning and the end until the portion of the sequence that is high quality is left between (the small vertical blue bars displayed in the mutation electropherogram). The portion of the sequence that is trimmed is considered to have a low signal to noise ratio.

These parameters and measures are summarised in the main graphical view (in a table below the electropherograms - as seen at the bottom of figure 6), however these values are not separately available in any of the tabulated report outputs.

3.7.2. Additional Mutation Surveyor outputs (tabulated output):

- *Size*; is the length of the sequence fragment (quality read length) after automated quality trimming of the terminal sequences (denoted by small vertical blue bars in the graphical output - figure 7).
- *The Lane Quality*; is a measure of the average signal to noise ratio defined by measuring the signal to noise ration of each nucleotide base and then take the average of the ratios. For example, a lane quality score of 20 signifies that there is 5% noise in that lane. ($s/n = 1/20 = 0.05$).

In the default analysis options (figure 14) the lane quality threshold is set to zero so lanes with a signal to noise ratio equal to zero are rejected and listed as 'Low Quality.' A call of 'Low Quality' usually signifies a high noise level, such that the software is unable to de-convolute and determine the underlying DNA sequence. Similarly, if there are a large number of N calls due to noise, then the lane quality will be low.

- *Total Score*; calculated value in the main graphical view (figure 6) - the sum of the mutation scores when a mutation is present in both forward and reverse traces. In other words in a bi-directional

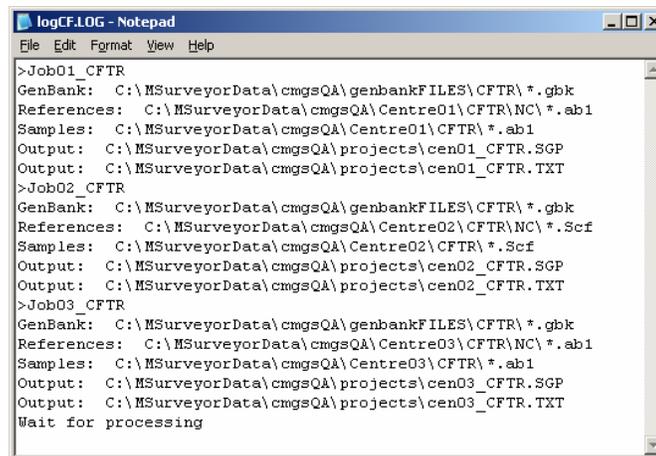
analysis, the total mutation score represents the sum of mutation scores from both sequence orientations, consequently the mutation calling thresholds are double those of a uni-directional analysis.

- *Number*; trace number by order of processing.
- *Sample File*; name of sample file.
- *Reference File*; name of the reference sequence file used in the analysis.
- *Dir*; indicates the sequencing orientation relative to the GenBank file.
- *Gene*; the gene name - feature taken from a *.seq or *.gbk
- *Exon*; the exon name - feature taken from a *.seq or *.gbk reference file.
- *RF*; the reading frame of the first base in an exon.
- *Start*; the number of the base at the start of processable data (indicated by a small vertical blue bar - figure 7).
- *End*; the number of the base at the end of processable data (indicated by a small vertical blue bar - figure 7).
- *Mut#*; the number of mutations found in this sample sequence. Where -1 indicates 'Bad Data'.
- *Output Mutation*; is the abbreviated name for each mutation in a sample given in the following order: base number, reference nucleotide, mutation nucleotide, and mutation score.

3.8. Analysis procedure adopted for this study

Each set of data to be tested was sub-divided into a manageable number of samples along with their respective wild type/normal control sequences (limited to a maximum of ~400 sequence traces for a single Mutation Surveyor run). These files were grouped into folders and a Mutation Surveyor autorun log was generated (figure 13).

This is a simple text file which can be generated using the log file generator within Mutation Surveyor or simply composed using a text editor, where the locations of all the folders / files used in the analysis are specified along with the desired location of the automated output files. Several jobs can be outlined in a single log file to aid batch processing of multiple projects.



```

logCF.LOG - Notepad
File Edit Format View Help
>Job01_CFTR
GenBank: C:\MSurveyorData\cmgsQA\genbankFILES\CFTR\*.gbk
References: C:\MSurveyorData\cmgsQA\Centre01\CFTR\NC\*.ab1
Samples: C:\MSurveyorData\cmgsQA\Centre01\CFTR\*.ab1
Output: C:\MSurveyorData\cmgsQA\projects\cen01_CFTR.SGP
Output: C:\MSurveyorData\cmgsQA\projects\cen01_CFTR.TXT
>Job02_CFTR
GenBank: C:\MSurveyorData\cmgsQA\genbankFILES\CFTR\*.gbk
References: C:\MSurveyorData\cmgsQA\Centre02\CFTR\NC\*.Scf
Samples: C:\MSurveyorData\cmgsQA\Centre02\CFTR\*.Scf
Output: C:\MSurveyorData\cmgsQA\projects\cen02_CFTR.SGP
Output: C:\MSurveyorData\cmgsQA\projects\cen02_CFTR.TXT
>Job03_CFTR
GenBank: C:\MSurveyorData\cmgsQA\genbankFILES\CFTR\*.gbk
References: C:\MSurveyorData\cmgsQA\Centre03\CFTR\NC\*.ab1
Samples: C:\MSurveyorData\cmgsQA\Centre03\CFTR\*.ab1
Output: C:\MSurveyorData\cmgsQA\projects\cen03_CFTR.SGP
Output: C:\MSurveyorData\cmgsQA\projects\cen03_CFTR.TXT
Wait for processing

```

Figure 13. Screen snapshot of an autorun log file.

The autorun feature is set to detect a log file in a user specified folder at user defined time intervals. When a log file awaiting processing is detected the Mutation Surveyor application is launched and the individual job folders interrogated for the files awaiting processing. Following automated mutation detection each processed autorun log file job produces its own results output file of the detected mutations along with a standard analysis project file. The result output file is a tab delimited simple text file of the advanced bi-directional report output.

The output mutation tables derived from the autoruns were then checked against the project file to confirm the identified mutations. For each separate data set, text output files were copied to an Excel file pending further analysis of the total data set.

The definition of automated mutation detection we use only includes mutations/variants included in the output mutation tables derived from the autorun feature, and so discounts any Mutation Surveyor feature that highlights variants within the graphical view of Mutation Surveyor and not be reported in the tabulated output.

3.9. Default analysis settings

The default 2 direction (bi-directional) analysis settings are claimed to be suitable for the majority of data and to have very low false positive and negative rates. In Mutation Surveyor these parameters are adjustable, whereas the Mutation Explorer version of the software is 'hard-wired' with the default setting which can not be altered by the user. Again the main mutation detection parameter thresholds are outlined below and shown in figure 14. Having taken a bi-directional sequencing data analysis strategy/approach for this study, note that we used the default settings which are pre-set for a bi-directional analysis with the appropriate score thresholds. Consequently we have not specifically attempted to assess the appropriateness of Mutation Surveyor to analyse single orientation sequence reads.

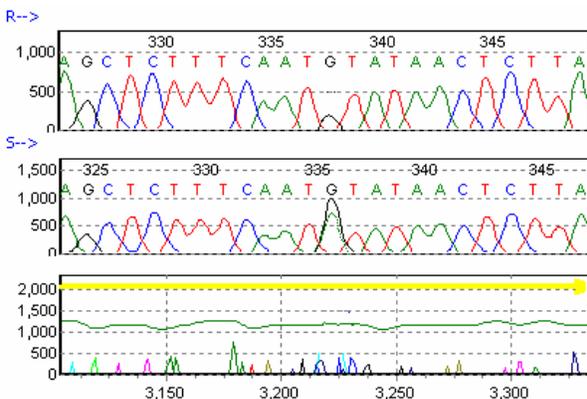
- *Mutation Height* (500) is the threshold height above which a peak in the mutation electropherogram is registered as a mutation
- *Overlapping Factor* (0.20) determines the minimum degree of overlap for a mutation to be registered.
- *Dropping Factor* (0.20) determines the minimum dropping factor for a mutation to be registered.
- *S/n Ratio* (1.00) determines how large the signal has to be relative to neighbouring noise in order for a mutation to be registered.
- *Mutation Score* (5.00) is used to call a mutation and rank its confidence level. Mutation score is a measure of the probability of error and is based on the S/n Ratio, overlapping factor and dropping factor.

3.10. Illogical data

In order to prevent the calling of false positive mutations, Mutation Surveyor has been designed to delete mutations that arise under certain circumstances - high background noise and what SoftGenetics call 'illogical mutations'.

False positive mutations can be identified in a noisy trace when both forward and reverse traces are used. Mutation Surveyor will automatically delete the mutations if the trace signal in a sequence read is less than three times the noise in the local section. Furthermore, if the dropping factor is <0.2, the mutation will be ignored, this intensity dropping factor is used to eliminate false positives due to sequencing instrument spikes.

Mutation Surveyor will also miss 'illogical mutations' - where the relative peak intensities (defined as the average of the four neighbouring peaks excluding the two peaks adjacent to the mutation peak) go against the rule that "the relative intensity of homozygous base is about 2 times higher than that of heterozygote for human DNA". The most common situation is when the relative intensity of a heterozygous base is greater than the intensity of the equivalent base in the homozygous normal control trace. Under normal circumstances, the relative intensity of a heterozygote sample is always lower than that of the homozygote.



A figure taken from the Mutation Surveyor manual (this is neither from our sample data nor real data; the data has been modified to demonstrate the principle). In this example the heterozygous relative peak height is 2 times higher than that of the homozygote reference, which is viewed as being illogical and hence missed by the automated mutation detection algorithm.

The figure displays six screenshots of the Mutation Surveyor software interface, arranged in a 2x3 grid. Each screenshot shows a different tab of the software's configuration options. Orange circles highlight specific parameters that were modified for the study.

- Top Left Screenshot:** Shows the 'Contig' tab. The 'BasePatch' checkbox is circled in orange.
- Top Middle Screenshot:** Shows the 'Mutation' tab. The 'CDS Only' checkbox in the 'Output Limits' section is circled in orange.
- Top Right Screenshot:** Shows the 'Output' tab. The 'Region of Interest Only' checkbox in the 'Output Limits' section is circled in orange.
- Bottom Left Screenshot:** Shows the 'Mutation Call Thresholds' tab. It displays a table of thresholds for 'Reject (<)' and 'Pass (>)' for various parameters.
- Bottom Middle Screenshot:** Shows the 'Settings' tab. It displays configuration for '2 Directions' and '1 Direction' with 'Reject' and 'Pass' thresholds.
- Bottom Right Screenshot:** Shows the 'Database URL' tab. It displays the 'Input the Database URL' checkbox and the 'User Local Directory' path.

Figure 14. Screen snapshots of Mutation Surveyor default analysis options [these settings were used throughout the course of this study otherwise stated, or with the exception of those parameters circled in orange - see section headed 'Analysis settings' for each data set]

4. PERFORMANCE TESTING OF MUTATION SURVEYOR

To test the performance of the automated mutation detection of Mutation Surveyor™ v2.51 we examined four sets of sequence data that cover a range of direct sequencing strategies from whole gene mutation scanning to single fragments/exons.

Two of these data sets (sections 4.1 and 4.2) are single sequence fragment data generated in a number of different diagnostic laboratories using different sequencing chemistries and platforms. The other two sets of data (sections 4.3 and 4.4) were generated in-house using two different methods; a VariantSEQr™ primer set (section 4.3) and Exon linked sequencing (section 4.4).

The data sets chosen for analysis in this study are intended to represent a spectrum of the type and quality of sequence data generated in clinical diagnostic laboratories.

Source of Data Sets

- 1 - European Molecular Genetics Quality Network (EMQN, see <http://www.emqn.org>) sequencing EQA scheme 2003 data set - a single sequence fragment from the CFTR (Cystic Fibrosis Transmembrane conductance Regulator) gene.
- 2 - Clinical Molecular Genetics Society (CMGS) study data set - a single sequence fragment from three different genes, the CFTR (Cystic Fibrosis Transmembrane conductance Regulator), Cx26 (Connexin 26, aka: Gap Junction Protein Beta-2; GJB2) and MEN1 (Multiple Endocrine Neoplasia, Type I) genes.
- 3 - VariantSEQr™ data set (PN 4327098)- 34 sequence fragments for the NF2 (Neurofibromatosis, Type II) gene
- 4 - Exon linked data set (Wallace *et al.* 2004) - 4 long-read sequence fragments from the NF2 gene.

Definitions of False results

- A *false positive* analysis was recorded for any unexpected mutation called and tabulated by Mutation Surveyor.
- Many of the false positive results were seen to be due to; noise, low signal intensity, loss of resolution or a sequencing artefact (e.g. a dye blob or spike) in either the test sample or the normal control sample. In such instances these false positives are then re-grouped as *poor quality sequence data*.
- A *bi-directional false negative* analysis was recorded for any mutation when Mutation Surveyor failed to detect the presence of an expected mutation in both orientations.
- A *uni-directional false negative* has been recorded where a mutation was not detected in one orientation although the mutation was detected in the sequence data from the opposite strand.
- Throughout this document all counts of false negative and positive results (in the summary tables 3, 7, 10 and 14) are expressed as the sum of single fragment failures, i.e. the numbers include both forward and reverse directions for bi-directional failures and all uni-directional failures from the given data set.

4.1. European Molecular Genetics Quality Network, DNA sequencing EQA scheme data

4.1.1. Data Source

EMQN External Quality Assessment scheme for DNA sequencing (2003). See www.emqn.org

4.1.2.1. Participants:

Sequence data from fifty one EMQN registered laboratories across Europe that participated in the external quality assessment scheme for DNA sequencing (2003). The services offered by participating laboratories focus predominantly on mutation detection. The majority of laboratories participating in the scheme were public (74%) and private (11%) sector diagnostic laboratories with the remaining number of labs offering research / diagnostic services or commercial sequencing.

4.1.2.2. Structure of the scheme:

Laboratories were provided with four test DNA samples plus a wild type control sample for sequencing. The samples constituted purified DNA that had been prepared by PCR from genomic DNA, which covered exon 10 of the CFTR gene. The test samples covered a range of different genotypes, all of which were likely to be encountered by laboratories in their day to day work.

Fragment	PCR Product size [bp]	Achieved Average Sequence Length [bp]	Number of Samples	TOTAL Length of Sequence Data [kb]
CFTR exon 10	460	440*	255	224.4
CFTR exon 10 ROI		221		112.7

Table 1. Calculated estimate of total length of sequence data analysed and Region Of Interest (ROI) for the whole data set (* = an average of the sequence size [defined as size in the Mutation Surveyor output table], calculated from 10 samples; ROI = coding sequence + 15bp either side of intronic sequence).

4.1.2.3. Instrumentation and sequencing chemistry

The type of sequencing platforms and sequencing chemistries used by the laboratories that participated in the EQA scheme is summarised in tables 1.1 and 1.2 respectively.

Sequencing Instrument	Number of Labs	Percentage of Total Number of Labs
Applied Biosystems ABI 3100	27	53%
Applied Biosystems ABI 3730	8	16%
Applied Biosystems ABI avant	4	8%
Applied Biosystems ABI 310	4	8%
Applied Biosystems ABI 377	3	6%
Applied Biosystems ABI 3700	1	2%
Beckman Coulter CEQ 2000	1	2%
Beckman Coulter CEQ 8000	1	2%
GE Healthcare MegaBase 500	1	2%
MJ Research BaseStation	1	2%
TOTAL	51	100%

Table 1.1. Sequencing Platforms / Instruments used by scheme participants

Sequencing Chemistry	Number of Labs	Percentage of Total Number of Labs
Applied Biosystems BigDye v3.1	25	49%
Applied Biosystems BigDye v1.1	18	35%
Applied Biosystems BigDye v2.0	2	4%
Applied Biosystems BigDye v3.0	2	4%
Beckman Coulter CEQ DTCS	2	4%
GE Healthcare DYEnamicET	2	4%
TOTAL	51	100%

Table 1.2. Sequencing Chemistries used by scheme participants

4.1.3. Analysis settings

The default Mutation Surveyor settings were used as outlined in section 3.9 (figure 14) with the option selected to display only those mutations detected within exons and 15bp of intronic sequences flanking the exon.

4.1.4. Timing and work load

Automated batch files were set up for data from each individual centre comprising the four test sample files to be compared to the wild type control trace file and a GenBank reference sequence file. Data for each centre generated its own text output file of the detected mutations along with an analysis project file (a compressed file of all the data and a processed output interface).

- Timing of automated analysis for data from a single centre was approximately 10 seconds.
- The output mutation table was then checked against the project file to confirm the identified mutations, this manual check took around 7 minutes.
- Total for 51 analyses (112.7Kb - based on ROI) = 6 man hours OR
- Average analysis time per 10Kb of sequence data = 31 minutes.

Sample Information				Mutation Surveyor Output Results							
Gene	Sample	Mutation	State	Expected Detection	Direction	Ave. Mutation Surveyor Quality Score	Mutations				
							Detected	Expected	Correct	False Negative	False Positive
CFTR	1	c.1519_1521delATC (p.Ile507del) [*1651_1653delATC]	heterozygous	delATC / 1408AG>A	F	31	102	102	100	2 [0]	2 [30]
					R	12	49	100	45	55 [12]	3 [0]
	2	c.1408A>G (p.Met470Val) [*1540A>G]	heterozygous	-	F	46	1	0	0	0	1 [47]
					R	38	1	0	0	0	1 [5]
	3	c.1408A>G / c.1477C>T (p.Met470Val / p.Gln493X) [*1540A>G / 1609C>T]	heterozygous	1477C>T	F	53	51	51	51	0	0
					R	43	51	51	49	2 [37]	2 [26]
	4	c.1408A>G / c.1466C>A (p.Met470Val / p.Ser489X) [*1540A>G / 1598C>A]	homozygous	1466C>A / 1408AG>G	F	52	103	102	101	1 [22]	2 [51]
					R	43	100	100	100	0	0
	Normal Control	c.1408A>G (p.Met470Val) [*1540A>G]	heterozygous	-	F	67	-	-	-	-	-
					R	47	-	-	-	-	-
TOTAL							458	506	446	60	11

Table 2. Summary information from the EMQN data set with observed and expected results [* = mutations cited using non-HGVS/historic CFTR naming, where nucleotide 1 is the first nucleotide of transcription - GenBank accession no. M28668; the numbers in the expected mutations column vary according to the number of successful PCR amplifications and/or sequencing reactions; false negative /positive values cited in square brackets denote the average Mutation Surveyor lane Quality score for those results]

4.1.5. Mutation Detection

No.	Sample File	Reference File	Dir	Qualit	Mut#	Mutation1	Mutation2	Mutation3	Mutation4
1	106_SEQ2003_1_F	106_SEQ2003_C_F_1-F	33	2	2	c.1408AG>A,p.MV470M\$30			c.1519_1521het_delATC
2	106_SEQ2003_1_R	106_SEQ2003_C_R_1-R	24	1	1	n.a.	n.a.	n.a.	c.1516_1518het_delATC
3	106_SEQ2003_2_F	106_SEQ2003_C_F_1-F	39	0	0				
4	106_SEQ2003_2_R	106_SEQ2003_C_R_1-R	35	0	0				
5	106_SEQ2003_3_F	106_SEQ2003_C_F_1-F	80	1	1			c.1477C>CT,p.Q493QX\$38	
6	106_SEQ2003_3_R	106_SEQ2003_C_R_1-R	66	1	1			c.1477C>CT,p.Q493QX\$36	
7	106_SEQ2003_4_F	106_SEQ2003_C_F_1-F	52	2	2	c.1408AG>G,p.MV470V\$26	c.1466C>A,p.S489X\$124		
8	106_SEQ2003_4_R	106_SEQ2003_C_R_1-R	59	2	2	c.1408AG>G,p.MV470V\$66	c.1466C>A,p.S489X\$146		
8				48	1.12	50.0%	25.0%	25.0%	25.0%

Figure 17. An example of the advanced bi-directional output table for data from a single centre [the mutations tabulated in the main body of the text correspond to those outlined in table 2]; note how the c.1408A>G mutation for the first sample is highlighted in red, as the mutation score lies below the threshold level as defined in the two direction analysis settings - see figure 14.

For this data set all but one (sample 3, mutation c.1477C>T) of the observed false positive results were seen to be due to poor quality sequence data (i.e. noise, low signal intensity, or loss of resolution) or a sequencing artefact (i.e. a dye blob or spike) in either the test sample or the normal control sample.

Summarised Result	Number of Mutations	Ave. Mutation Surveyor Quality Score
Detected mutations	458	-
Expected mutations per strand sequenced	506	-
Correctly identified mutations	446	40
False positives	11 (2.2 %)	29
False negatives	60 (11.9%)	13

Table 3. Summarised Results [% values in brackets = false negative and positive rates expressed as a percentage of the expected number of mutations per strand sequenced]

4.1.5.1. False negatives broken down by directional coverage

Bi/uni-directional	Explanation on Visual Inspection	Number of Negatives
Bi-directional false negatives	<i>due to poor quality data</i>	2
	<i>true bi-directional false negative</i>	0 (0%)
Total		2
Uni-directional false negatives	<i>due to poor quality data</i>	6
	<i>mutation masked by frameshift (sample 1R)</i>	47
	<i>true uni-directional false negative</i>	3 (0.6%)
Total		56

Table 4. Breakdown of false negative results [see section 4.1.5.3. Screen snapshots; % values in brackets = true false negative rates expressed as a percentage of the total expected number of mutations per strand sequenced for this data set]

4.1.5.2. Comments

On further examination of the samples with true false positive/negative results there was no clear correlation with the sequencing instrumentation and chemistry (these are indicated on each illustration below - Section 4.1.5.).

All mutation calls were made correctly except those for the mutation c.1519_1521delATC, (p.Ile507del). In the forward orientation the majority of the expected mutations are called correctly (46/50), whereas in the reverse orientation most of the samples were consistently named incorrectly as c.1516_1518het_delCAT (41/45) with only one sample being named correctly (table 4.1).

Although the mutation c.1519_1521delATC, p.Ile507del is an in-frame deletion, the protein change is not interpreted for this mutation by Mutation Surveyor.

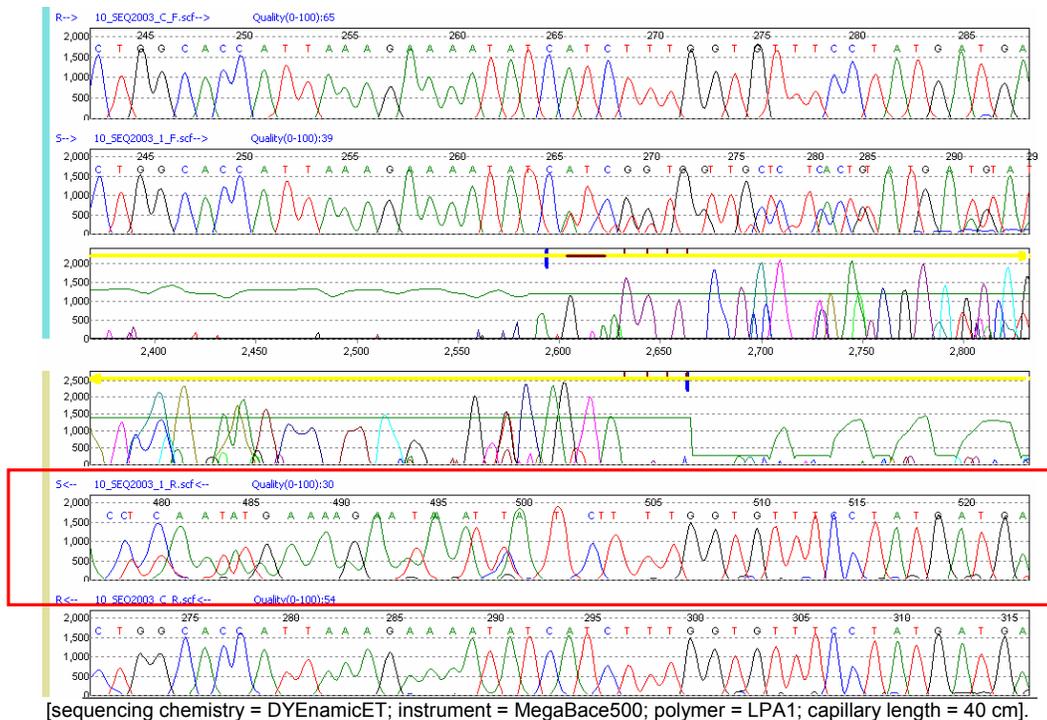
Furthermore sample 1 which was heterozygous for c.1519_1521delATC also has another difference from the control sequence (c.1408AG>A) which was expected to be detected by the software. Mutation Surveyor consistently failed to identify this mutation downstream of the frameshift mutation although Mutation Surveyor is claimed to de-convolute the sequence into both alleles and permit continuation of mutational analysis downstream of an indel mutation.

Strand	Correct Name	Incorrect Name	Number
Forward	c.1519_1521delATC		46
		c.1529_1531het_delTTT	1
		c.1515_1517het_delTAT	1
		c.1516_1518het_delATC	1
		c.1515_1517het_delTAT	1
Reverse	c.1519_1521delATC		1
		c.1516_1518het_delATC	41
		c.1518_1520het_delCAT	1
		c.1520_1522het_delTCT	1
		c.1523_1525het_delTTG	1

Table 4.1. Breakdown of correct and incorrect naming of mutation c.1519_1521delATC

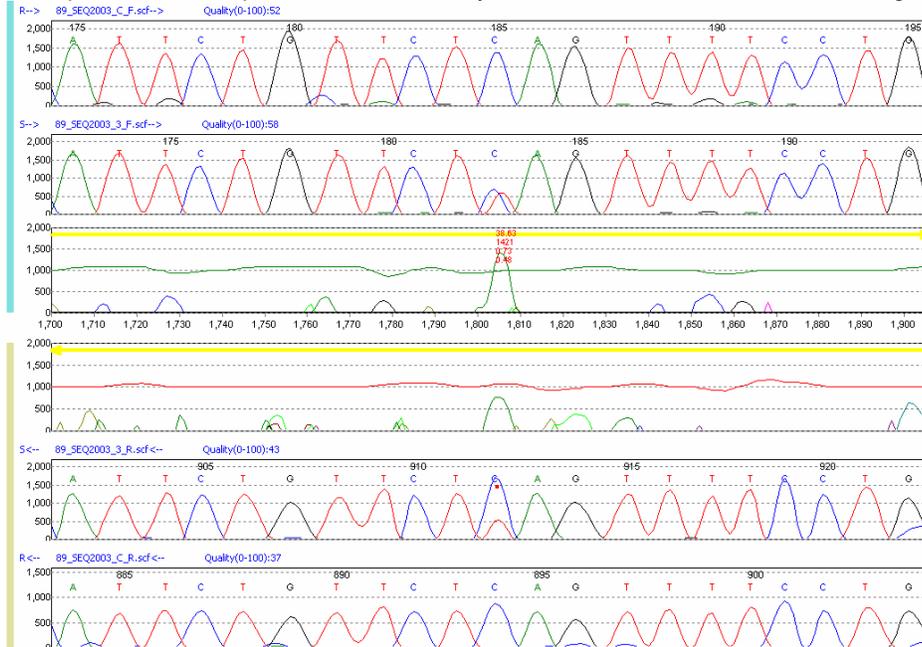
4.1.5.3. Screen snapshots true of uni-directional false negatives:

- i) Frameshift undetected in reverse direction (boxed in red, EMQN sample 1, mutation c.1519_1521delATC). According to SoftGenetics the basecalling software employed by this instrument (MegaBace500) stretches data for heterozygous indel traces. Therefore a mixture of two peaks in a location is stretched to two individual bases and then called as two bases with a large migration time stretch. In this snapshot the peak overlap information is lost and may result in one of two problems - misalignment of data or failure to detect indels.



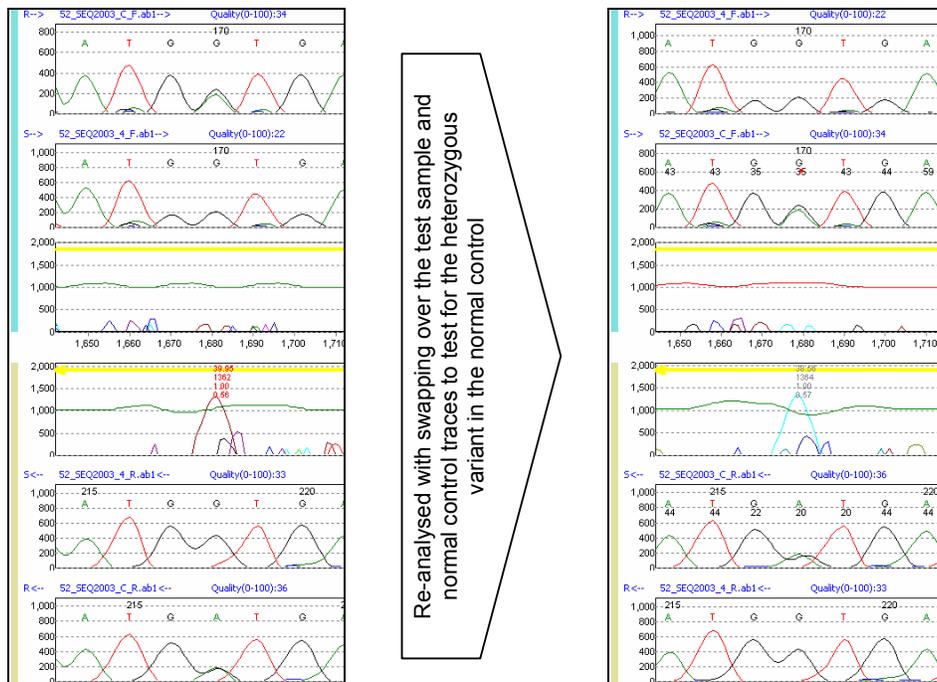
SoftGenetics recognise both of the following false negative examples as being 'illogical data', and that they are due to data from different instruments/chemistries or different run conditions. However following closer examination of the data files, to the best of our knowledge this data was generated using consistent experimental conditions. The only discrepancy noted was of that for example (iii) where the two samples (test and control) were run one day apart on the sequencer and this may have introduced some variability from the running conditions.

- ii) Undetected substitution in reverse direction (EMQN sample 3, mutation c.1477C>T). The relative intensity of C in the wild type trace is lower than that of the mutant C. *N.B* a red dot has been placed above the missed variant to signify peaks where the relative intensity drops 35% but the other mutation requirements are not met and that these points marked by red dots should be reviewed carefully, nevertheless the software does not highlight these in any of the tabulated outputs and so it is up to the user to manually scan the traces and look out for these highlights.



[sequencing chemistry = DYEnamicET; instrument = BaseStation; polymer = kilobase pack; capillary length = 30 cm]

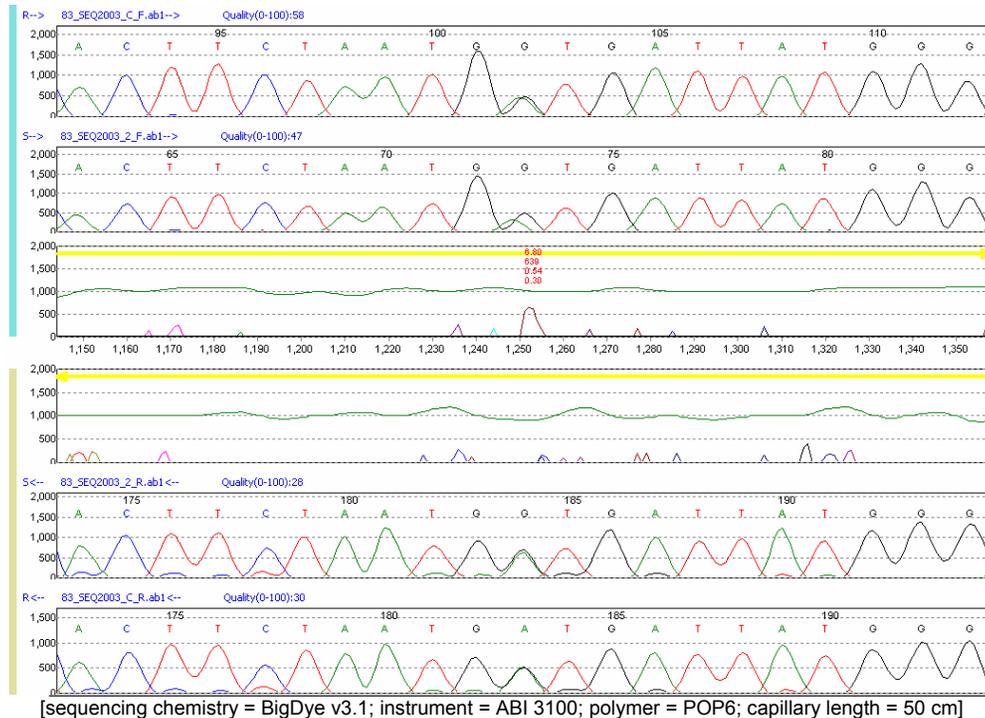
- iii) Undetected substitution in forward direction (EMQN sample 4, mutation c.1408AG>G). The relative intensity of G in the wild type trace is higher than the mutant G in the GA heterozygote. In addition SoftGenetics recommend using a heterozygote reference. So we decided to swap over the samples and assign the normal control as the test sample and *vice-versa*, however this resulted in no automated mutation detection, as the reverse variant is now deleted by the software (denoted by grey type above the mutation peak) and the forward strand variant now has a red dot above the variant:



[sequencing chemistry = DigDye v3.0; instrument = ABI 377; polymer = 5% PAGE; plate length = 36 cm]

4.1.5.4. Screenshot of true uni-directional false positive:

- i) All the false positives observed within this data set were due to poor sequence data or a sequencing artefact in either the test sample or the normal control trace. However in this instance the mutant peak is slightly shifted from the wild type peak and is called as a mutation even though both test and control sample were heterozygous for the same polymorphism. SoftGenetics say this has occurred because the reference is a heterozygote at this position and that they do not recommend using a heterozygote reference:



4.2. UK Clinical Molecular Genetics Society (CMGS), comparative study of diagnostic sequencing data

4.2.1. Data Source

UK CMGS comparative assessment of sequencing data generated by diagnostic laboratories (2003).

4.2.1.1. Participants:

Twelve CMGS laboratories throughout the UK that participated in a comparative study of diagnostic sequencing data (2003). All the participating centres are public sector NHS laboratories that focus on clinical mutation detection. One laboratory participated twice, running the sequencing data on two different platforms (ABI377 and ABI3100).

4.2.1.2. Structure of the scheme:

Each participating laboratory was provided with five test DNA samples plus a wild type control sample for a single exon region of three different genes that are tested in UK clinical molecular genetics laboratories: Cystic Fibrosis Transmembrane conductance Regulator (CFTR) exon 3; Connexin 26 (Cx26) exon 2, *aka*: Gap Junction Protein Beta-2, GJB2; and Multiple Endocrine Neoplasia, Type I (MEN1) exon 3. The samples were chosen to represent a range of different sequence changes that are encountered by diagnostic laboratories (see table 6 for full details). Each centre was required to produce bi-directional sequence data for eighteen different DNA samples (total of 234 samples sequenced for the whole data set).

Fragment	PCR Product size [bp]	Achieved Ave. Sequence Length [bp]	Number of Samples	TOTAL Length of Sequence Data [kb]
CFTR exon 3	308	270*	78	42.1
CFTR exon 3 ROI		138		21.5
Cx26 exon 2 / ROI	416	380*		59.3
MEN1 exon 3	377	335*		52.3
MEN1 exon 3 ROI		239		37.3
Whole data set	1101	985	234	153.7
Whole data set ROI		757		118.1

Table 5. Calculated estimate of total length of sequence data analysed and Region Of Interest (ROI) for the whole data set (* = an average of the sequence size [defined as size in the Mutation Surveyor output table], calculated from 10 samples; ROI = coding sequence + 15bp either side of intronic sequence. *N.B* PCR primers for the Cx26 exon 2 are within the exonic sequence and so the entire sequence product is composed of the ROI).

4.2.2.3. Instrumentation and sequencing chemistry

The type of sequencing platform and sequencing chemistry used by the laboratories that participated in the study is summarised in tables 5.1 and 5.2 respectively.

Sequencing Instrument	Number of Instruments	Percentage of Total Number of Instruments
Applied Biosystems ABI 3100	7	54%
Beckman Coulter CEQ 8000	4	31%
Applied Biosystems ABI 310	1	8%
Applied Biosystems ABI 377	1	8%
TOTAL	13	100%

Table 5.1. Sequencing Platforms/Instruments used by scheme participants

Sequencing Chemistry	Number of Labs	Percentage of Total Number of Labs
Applied Biosystems BigDye v1.1	4	31%
Beckman Coulter CEQ DTCS	4	31%
Applied Biosystems BigDye v2.0	3	23%
Applied Biosystems BigDye v3.1	2	15%
TOTAL	13 *	100%

Table 5.2. Sequencing Chemistries used by scheme participants
[* = one of the participating laboratories performed the test/ran the samples on two different platforms]

Sample Information				Mutation Surveyor Output Results						
Gene	Sample	Mutation	State	Direction	Ave. Mutation Surveyor Fragment Quality	Mutations				
						Detected	Expected	Correct	False Negative	False Positive
CFTR	1	c.200C>CT (p.Pro67Leu) [*332C>CT]	heterozygous	F	46	18	13	11	2 [15]	7 [34]
				R	53	24	12	11	1 [29]	13 [37]
	2	c.262_263delTT [*394_395delTT]	heterozygous	F	29	18	12	12	0	6 [25]
				R	16	10	10	9	1 [<0]	1 [0]
	3	c.254G>A (p.Gly85Glu) [*386G>A]	homozygous	F	46	17	13	12	1 [1]	5 [33]
				R	50	20	12	11	1 [24]	9 [43]
	4	c.221G>AG (p.Arg74Gly) [*353G>AG]	heterozygous	F	46	16	12	12	0	4 [53]
				R	47	21	13	12	1 [27]	9 [47]
5	c.178G>GT (p.Glu60X) [*310G>GT]	heterozygous	F	54	19	12	12	0	7 [44]	
			R	46	21	13	11	2 [14]	10 [41]	
Normal Control	-	-	F	63	-	-	-	-	-	
			R	66	-	-	-	-	-	
TOTAL	-	-	-	-	-	184	122	113	9 [14]	71 [36]
Cx26	1	c.368C>CT (p.Gln124X)	heterozygous	F	40	17	13	12	1	5 [35]
				R	41	17	13	13	0	4 [32]
	2	c.539A>AG (p.Glu114Gly)	heterozygous	F	38	19	13	13	0	6 [31]
				R	43	14	13	12	1 [9]	2 [56]
	3	c.436C>CT (p.Gln80X)	heterozygous	F	39	11	12	9	3 [45]	2 [26]
				R	44	13	12	12	0	1 [58]
	4	c.447C>CG (p.Phe83Leu)	heterozygous	F	38	13	13	10	3 [37]	3 [40]
				R	41	15	13	13	0	2 [15]
5	c.655G>AG (p.Val153Ile)	heterozygous	F	40	13	12	12	0	1 [19]	
			R	43	14	13	12	1 [17]	2 [41]	
Normal Control	-	-	F	47	-	-	-	-	-	
			R	49	-	-	-	-	-	
TOTAL	-	-	-	-	-	146	127	118	9 [35]	28 [35]
MEN1	1	c.604G>AG (p.Cys165Tyr)	heterozygous	F	36	13	10	9	1 [23]	4 [44]
				R	38	20	10	10	0	10 [28]
	2	c.658G>CG (p.Trp183Ser)	heterozygous	F	36	21	13	12	1 [22]	9 [33]
				R	36	17	13	12	1 [38]	5 [39]
	3	c.622G>AG (p.Arg171Gln)	heterozygous	F	38	10	12	10	2 [5]	0
				R	40	10	12	9	3 [9]	1 [11]
	4	c.681G>GT (p.Glu191X)	heterozygous	F	39	16	12	12	0	4 [9]
				R	41	18	12	12	0	6 [28]
5	c.571G>GT (p.Ser154Ile)	heterozygous	F	37	16	12	9	3 [22]	7 [34]	
			R	36	21	12	9	3 [25]	12 [17]	
Normal Control	-	-	F	61	-	-	-	-	-	
			R	52	-	-	-	-	-	
TOTAL	-	-	-	-	-	162	118	104	14 [19]	58 [29]

Table 6. Summary information with expected and observed results of the CMGS data set [* = mutations cited using non-HGVS/historic CFTR naming, where nucleotide 1 is the first nucleotide of transcription - GenBank accession no. M28668; the numbers in the expected mutations column vary according to the number of successful PCR amplifications and/or sequencing reactions; false negative /positive values cited in square brackets denote the average Mutation Surveyor lane Quality score for those results].

4.2.3. Analysis settings

Default settings were used as outlined in section 3.9 - figure 14.

4.2.4. Timing and work load

Automated batch files were set up for each of the genes from each individual centre comprising five test sample files to be compared to a wild type control trace file and a GenBank reference sequence file. Data for each gene and centre generated its own text output file of the detected mutations along with an analysis project file.

- Timing of automated analysis for data for a gene from a single centre was approximately 10 seconds.
- The output mutation table was then checked against the project file to confirm the identified mutations, this manual check took around 7 minutes.
- Total for 39 analyses (118.1Kb - based on ROI) = 4.55 man hours OR
- Average analysis time per 10Kb of sequence data = 23 minutes

4.2.5. Mutation Detection

Almost all of the observed false positive results were once again due to poor quality sequence data or a sequencing artefact in either the test sample or the normal control sample and no particular pattern of false positives were observed.

However the majority of the false positive and negative results that could not be explained by poor quality data clustered around those samples sequenced using the Beckman Coulter CEQ 8000 sequencer in combination with the CEQ DTCS sequencing chemistry. For this reason this data is shown as a separate category within table 8. The only other correlation observed with regards to the sequencing chemistries and instrumentation was that all of the true uni-directional false negative data were from the same centre (section 4.2.5.3.). Apart from these two observations within this data set there are no other significant failure patterns to be noted in this study when looking at the Mutation Surveyor software and its ability to deal with different sequencing chemistries and platforms.

Summarised Results	Number of Mutations	Ave. Mutation Surveyor Quality Score
Detected mutations	492	-
Expected mutations per strand sequenced	367	-
Correctly identified mutations	335	38
False positives	157 (42.8%)	33
False negatives	32 (8.7%)	23

Table 7. Summarised Results [% values in brackets = false negative and positive rates expressed as a percentage of the expected number of mutations per strand sequenced]

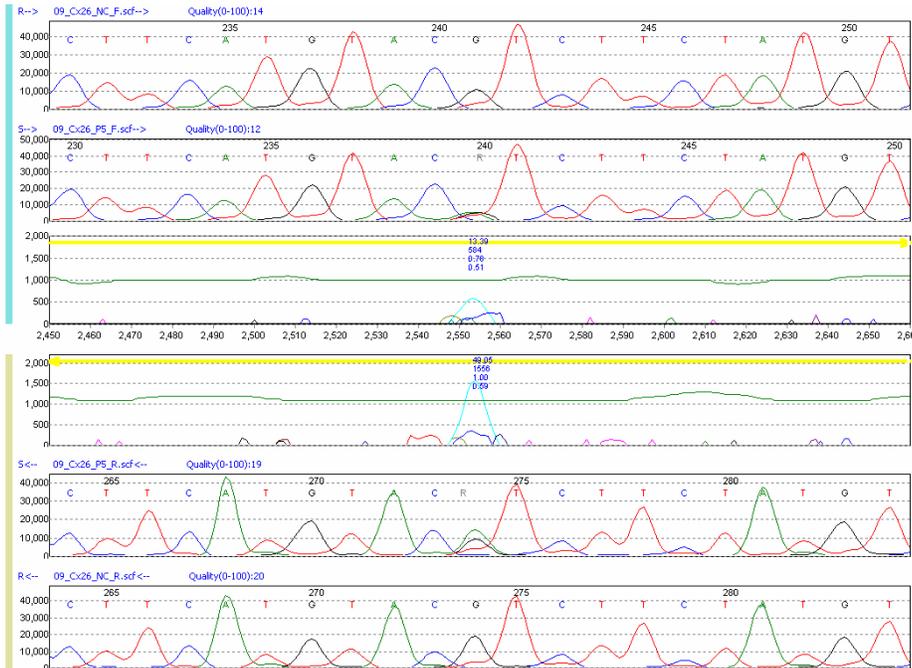
4.2.5.1. False negatives broken down by directional coverage

Bi/uni-directional	Explanation on Visual Inspection	Number of Negatives			
		CFTR	Cx26	MEN1	Total
Bi-directional false negatives	<i>due to poor quality data</i>	0	0	1	1
	<i>true bi-directional false negative (Beckman CEQ)</i>	0	0	0	0
	<i>true bi-directional false negative (other platforms)</i>	0	1	1	2 (1.1%)
Total		0	1	2	3
Uni-directional false negatives	<i>due to poor quality data</i>	2	2	2	6
	<i>mis-aligned data</i>	0	2	0	2
	<i>true uni-directional false negative (Beckman CEQ)</i>	7	5	6	18
	<i>true uni-directional false negative (other platforms)</i>	0	0	1	1 (0.2%)
Total		9	9	9	27

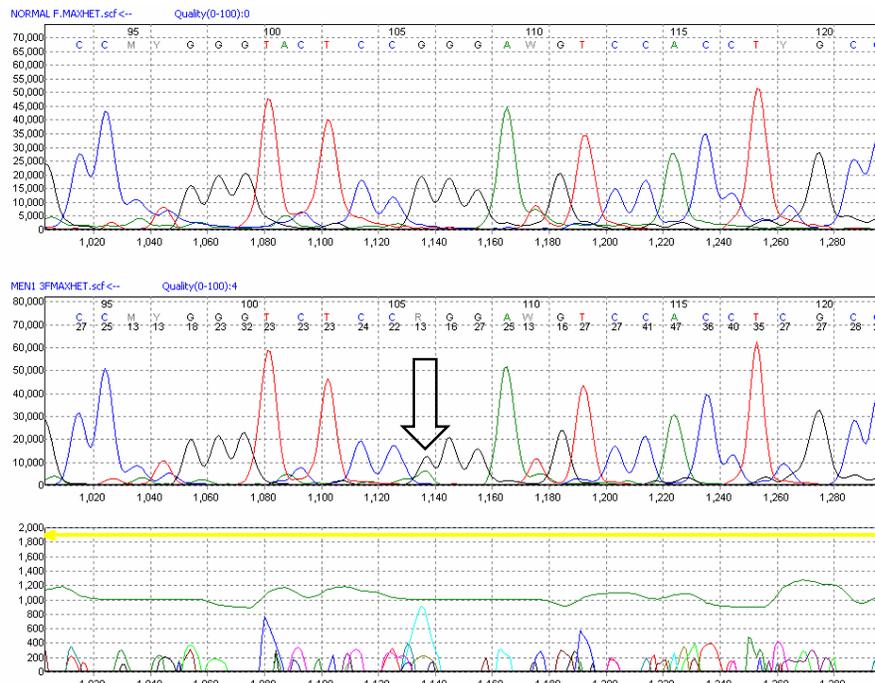
Table 8. Breakdown of false negative results [% values in brackets = true false negative rates expressed as a percentage of the total expected number of mutations per strand sequenced for this data set]

4.2.5.2. Comments

The majority of the uni-directional false negative results (18/27) are in data produced by centres using the Beckman Coulter CEQ 8000 instrument and the CEQ DTCS sequencing chemistry. An example of a good Beckman trace (all the data from this centre for the Cx26 gene was of good quality, where all expected mutations are observed and there are no false results) can be seen below, followed by an example of a false negative for the CMGS MEN1 sample 3 mutation c.622G>AG, indicated by the black arrow. On closer examination of the sequence trace morphology there is far greater variability in relative peak heights than the other sequencing chemistries, this may be affecting the analysis of the data in Mutation Surveyor. SoftGenetics do acknowledge that Beckman Coulter dye chemistry gives more variation in relative peak intensities and has more noise and as a result there maybe a greater number of false negatives for such data.



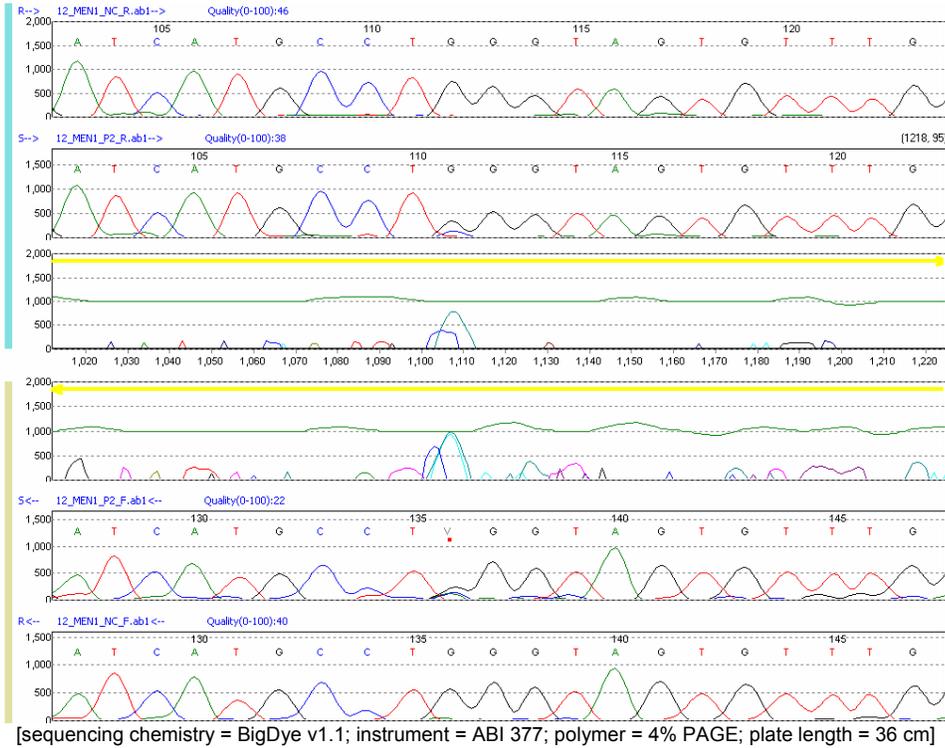
[sequencing chemistry = CEQ DTCS Quickstart; instrument = Beckman CEQ 8000; polymer = LPA1; capillary length = 33 cm]



[sequencing chemistry = CEQ DTCS Quickstart; instrument = Beckman CEQ 8000; polymer = LPA1; capillary length = 33 cm]

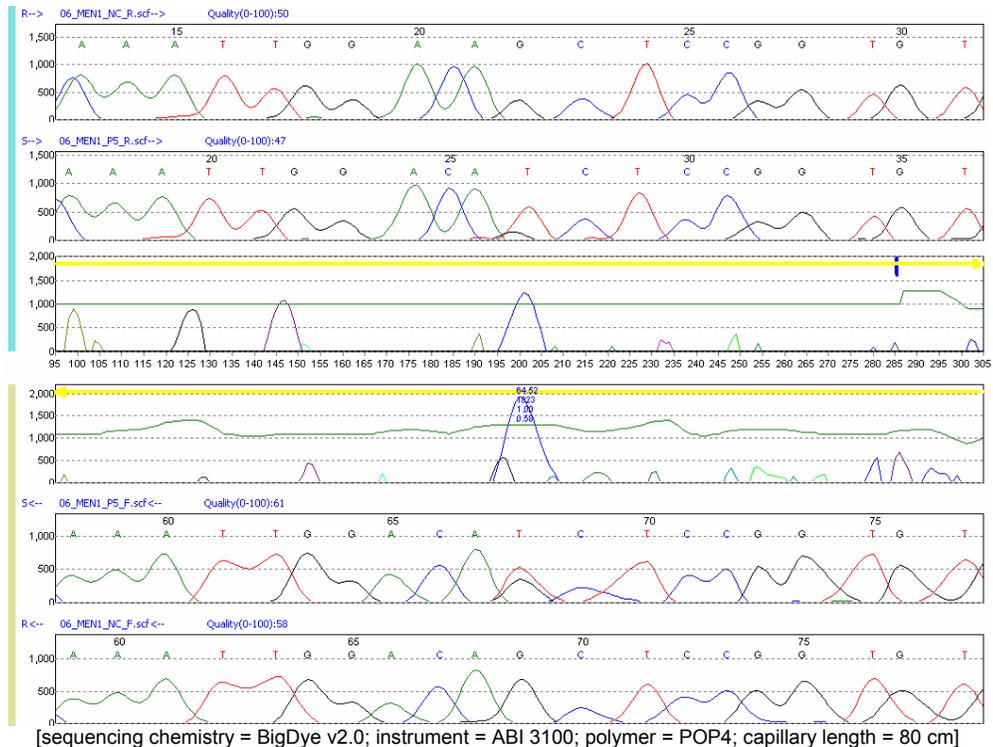
4.2.5.3. Screen snapshot of true bi-directional false negatives:

- i) A bi-directional false negative, there is some noise in both orientations but mutations are visible by eye (CMGS MEN1 sample 2, mutation c.658G>C; see section 4.1.5.3. ii, for a comment on the red dot seen above the missed mutation). SoftGenetics acknowledge that this problem is due to noise in the reverse and a small dropping factor in the forward and reverse directions:

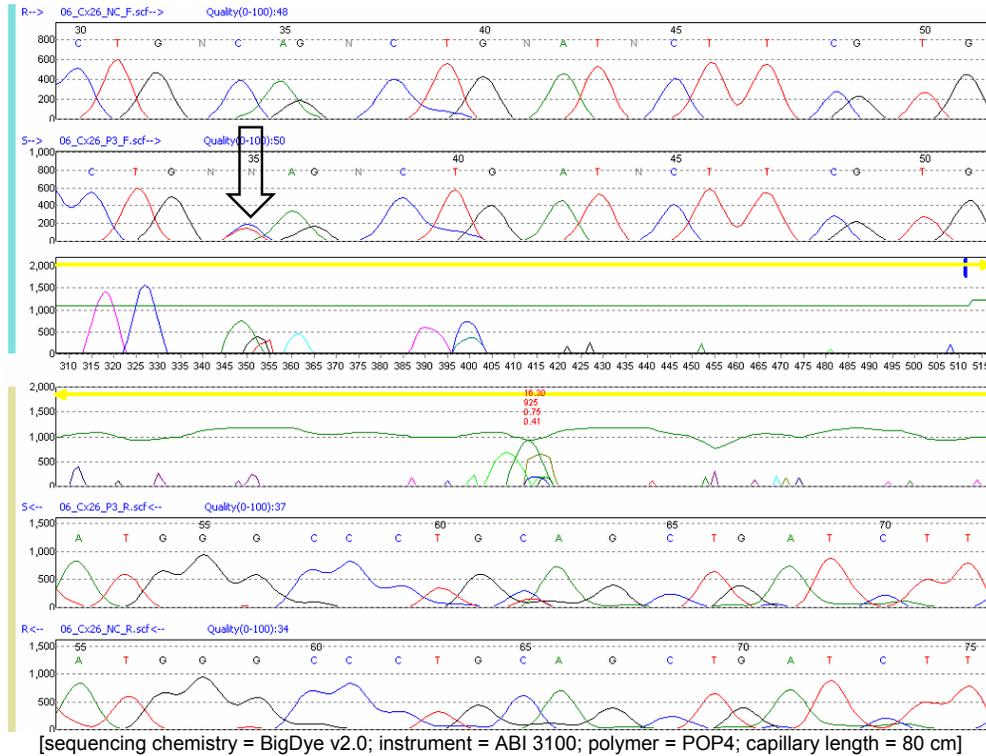


4.2.5.4. Screen snapshots of uni-directional false negatives:

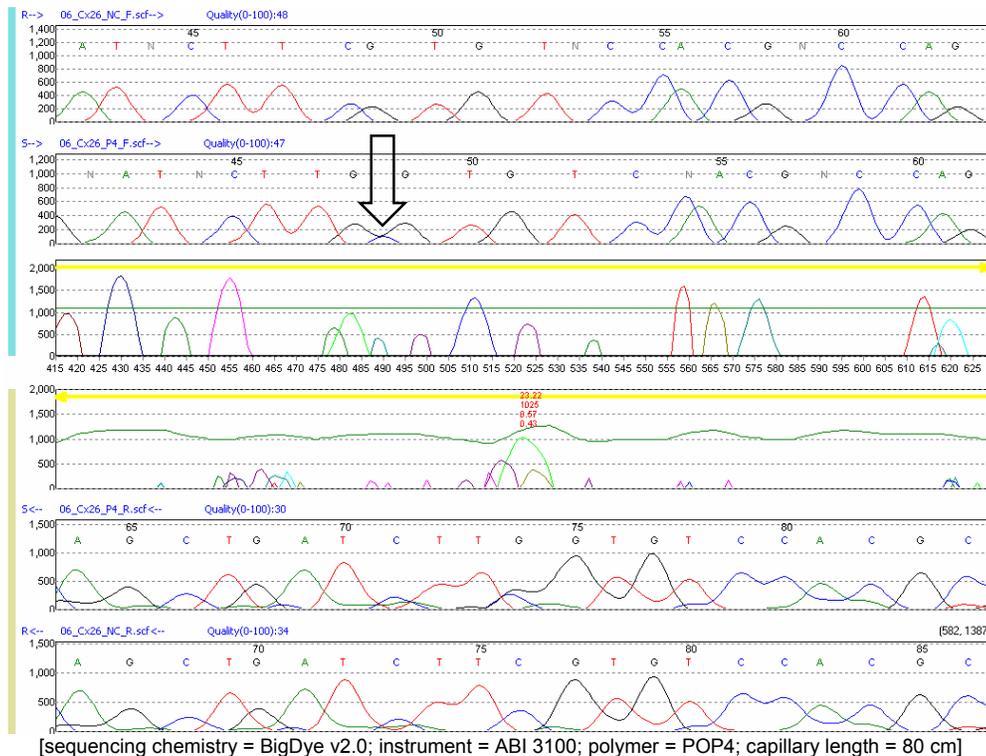
- i) CMGS MEN1 sample 5, mutation c.571G>T is missed in the F direction. SoftGenetics say this is a mosaic peak with a low dropping factor and that this is a problem with Mutation Surveyor they are planning to improve:



- ii) CMGS Cx26 sample 3, mutation c.436C>T, missed in F direction (indicated by the black arrow) also note how the F/R strands are misaligned. SoftGenetics note that the misalignment in these two samples is due to basecalling errors and that higher quality sequence should be obtained, however the software has accepted the trace data and given each trace an acceptable Quality score:



- iii) CMGS Cx26 sample 4, mutation c.447C>G, missed in the F direction (indicated by the black arrow) again note how the F/R strands are misaligned (see comment noted above):



4.3. VariantSeqr™ resequencing data set

4.3.1. Data Source

A panel of seventy six patients were sequenced for the entire neurofibromatosis type 2 (NF2 - OMIM #101000) gene region. Data was generated using the Applied Biosystem's NF2 gene VariantSeqr™ primer set (complete gene region including 5' and 3' UTRs; product code RSS00013418_02) comprising 34 amplimers. This data set was known to have 19 pathogenic sequence variants (table 12) as all the patients from this panel had previously been tested for mutations in the NF2 gene by direct sequencing (using our in-house exon linked/Meta-PCR - Wallace *et al.* 2004). The sequencing data used was the same data set analysed in our Technology Assessment Report, Mutation scanning - Applied Biosystems VariantSeqr™ and SeqScape v2.1®: an assessment using a model system January 2005.

4.3.1.1. Composition and coverage of the NF2 VariantSeqr™ kit data

- 34 fragments averaging 530bp in size cover the NF2 gene region of interest (this includes the 5' and 3' UTRs).
- 18 of the 34 fragments cover the coding region of interest (exons 1 to 17).
- The remaining 16 fragments cover 1.4kb of sequence upstream (5' UTR) and 3.8kb of sequence downstream (3' UTR) of the NF2 coding sequence.
- Coverage of intronic sequences flanking the exonic sequence by between 50 and 120bp.
- The majority of the sequencing fragments are overlapped by other fragments in the kit.
- Not all the data generated is bi-directional, 8/34 fragments produce usable sequence data in only one direction. 14/19 of the mutations in this data set have been sequenced in both F/R directions and 5 of them only in one direction.

Fragments / Region	Number of Patients	TOTAL Length of Sequence Data
All 34 fragments	1	31.8kb
ROI only		4.5kb
Whole data set	76	2.42Mb
Whole data set ROI		342kb

Table 9. Calculated estimate of total length of sequence data analysed and Region Of Interest (ROI) for the whole data set (ROI = coding sequence + 15bp either side of intronic sequence).

A more comprehensive breakdown and analysis of the sequencing data from this data set can be found in the NGRL (Manchester) Health Technology Assessment of the NF2 VariantSeqr™ primer set and the SeqScape® mutation analysis software from Applied Biosystems (http://www.ngrl.org.uk/Manchester/Pages/Downloads/SeqScapeHTA/Vseqr_SeqS_v5.pdf)

4.3.1.2. Instrumentation and sequencing chemistry

All of this data was sequenced using an ABI3730 capillary sequencer and BigDye v3.1 chemistry in strict adherence to the manufacturer's guidelines.

4.3.3. Analysis settings

Once again default settings were used as outlined in section 3.9 - figure 14, with the option selected to display only those mutations detected within exons and 15bp of flanking intronic sequences.

4.3.4. Timing and work load

This data set was broken up into batches (n=13) of around six test samples each (as this equated to the maximum number of samples a single job/project can handle at once ≤ 400 traces). Batches were prepared for automated analysis with a set of wild type sequence trace files and a NF2 GenBank file (based on NCBI35:22:28324117:28419 - see section 5.1. *Contig Alignment, Reference Sequences and Automated Mutation Naming in Mutation Surveyor*). As with the previous data sets each batch generated its own text output file of the detected mutations along with an analysis project file.

- Timing of automated analysis for data from a single batch of samples was approximately 8 minutes.
- The output mutation table was then checked against the project file to confirm the identified mutations, this manual check took around 5-10 minutes depending on the number of mutations reported by the software.
- Total for 13 batches (342Kb - based on ROI) = 2.2 man hours OR
- Average analysis time per 10Kb of sequence data = 3.8 minutes.

4.3.5. Mutation Detection

Summarised Results	Number of Mutations	Ave. Mutation Surveyor Quality Score
Detected mutations	249	-
Expected mutations per strand sequenced	41	-
Correctly identified mutations	37	28
False positives	212 * (517%)	20
False negatives	4 (9.7%)	24

Table 10. Results and Quality Score Summary. [* = all false positive results were observed to be due to poor quality sequence data; % values in brackets = false negative and positive rates expressed as a percentage of the expected number of mutations per strand sequenced]

4.3.5.1. False negatives broken down by directional coverage

Bi/uni-directional	Explanation on Visual Inspection	Number of Negatives
Bi-directional false negatives	<i>true bi-directional false negative</i>	1 (4.9%)
Total		1
Uni-directional false negatives	<i>mutation masked by frameshift in cis (sample S05)</i>	1
	<i>true uni-directional false negative</i>	1 (2.4%)
Total		2

Table 11. Breakdown of false negative results [% values in brackets = true false negative rates expressed as a percentage of the total expected number of mutations per strand sequenced for this data set]

4.3.5.2. Comments

As with the previous sections the majority of the false positive results were due to poor quality sequence data or a sequencing artefact in either the test sample or the normal control sample.

Over 90% of the positive control mutations were correctly identified in both directions as seen in table 10 above, the only exceptions to this were one positive control mutation missed in both directions and two samples where the mutation was only detected in a single direction.

One of the uni-directional missed mutations (data not shown) is a substitution (G>A) in sample S05 fragment 13 F; mutation c.810G>A,p.E270E, this is a false negative due to it being masked in the one sequence direction by a second compound frameshift mutation (Mutation Surveyor is claimed to continue mutation detection through a frame shifted sequence). Both the other false negatives are noted in the screen snapshots below Section 4.3.5.3.

Sample Information					Mutation Surveyor tabulated output				
Sample	Mutation	State	Fragment	Dir	Quality Score	Mutation1	Mutation2	Mutation3	Mutation4
S19	c.115-2 A>C	mosaic	4	F	33	IVS115-3A>ACS12			
				R	29	IVS115-3A>AC\$15			
			6	F	25				
				R	35				
S43	c.169C>T (p.Arg57X)	heterozygous	4	F	54		c.169C>CT,p.R57RX\$65		
				R	41		c.169C>CT,p.R57RX\$43		
			6	F	29		c.169C>CT,p.R57RX\$52		
				R	32		c.169C>CT,p.R57RX\$27		
S45	c.240+1G>C	heterozygous	4	F	29				IVS240+1G>CG\$45
				R	16				IVS240+1G>CG\$59
			6	F	26				IVS240+1G>CG\$34
				R	26				IVS240+1G>CG\$34
S51	c.203_206delTCAA*	heterozygous	4	F	32			c.203_206het delTCAA	n.a.
				R	51	n.a.	n.a.	c.200_203het delCAAT	
			6	F	29			c.203_206het delTCAA	n.a.
				R	39	n.a.	n.a.	c.200_203het delCAAT	
S05	c.737delC* + c.810G>A	mosaic	13	F	15	c.737het delC			
				R	21	c.735het delC			c.810G>AG,p.E270EE\$37
S42	c.784C>T (p.Arg262X)	heterozygous	13	F	22		c.784C>CT,p.R262RX\$40		
				R	26		c.784C>CT,p.R262RX\$41		
S57	c.784C>T (p.Arg262X)	heterozygous	13	F	38		c.784C>CT,p.R262RX\$43		
				R	26		c.784C>CT,p.R262RX\$45		
S38	c.1021C>T (p.Arg341X)	heterozygous	16	F	37	c.1021C>CT,p.R341RX\$77		n.a.	n.a.
				R	25	c.1021C>CT,p.R341RX\$36		n.a.	n.a.
S59	c.1446+2_1446+3delTG	heterozygous	18	F	22	IVS1446+4_1446+5het delAG	n.a.	n.a.	n.a.
				R	0	n.a.	n.a.	n.a.	n.a.
S61	c.288_290delCTT	heterozygous	8	F	22	c.288_290het delCTT	n.a.	n.a.	n.a.
				R	24	c.291_293het delCTT		n.a.	n.a.
S41	c.447+2T>C	heterozygous	9	R	14	IVS447+2T>CT\$100		n.a.	n.a.
S56	c.447+2T>C	heterozygous	9	R	38	IVS447+2T>CT\$81		n.a.	n.a.
S03	c.516+2T>C	heterozygous	10	F	24		IVS516+2T>CT\$26		n.a.
				R	37		IVS516+2T>CT\$144		n.a.
S50	c.459C>G (p.Tyr153X)	heterozygous	10	F	23	c.459C>CG,p.Y153YX\$56			n.a.
				R	41	c.459C>CG,p.Y153YX\$47			n.a.
S06	c.586C>T (p.Arg196X)	mosaic	11	F	25	c.586C>CT,p.R196RX\$26			n.a.
S37	c.586C>T (p.Arg196X)	heterozygous	11	F	25	c.586C>CT,p.R196RX\$31			n.a.
S53	c.592C>T (p.Arg198X)	heterozygous	11	F	12		c.592C>CT,p.R198RX\$47		n.a.
S58	c.1640delA	heterozygous	20	F	0	c.1640het delA	n.a.	n.a.	n.a.
				R	26	c.1640het delA		n.a.	n.a.

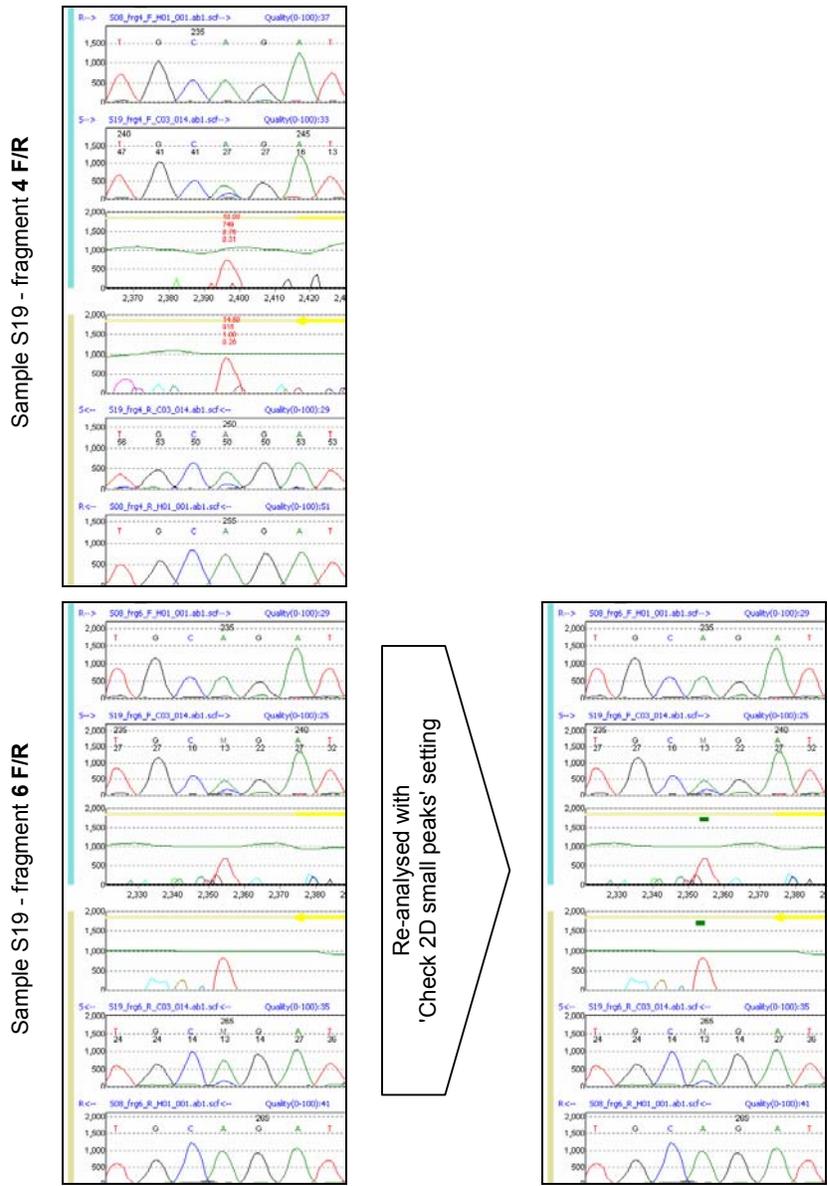
Table 12. Summary information with expected and observed results of the VariantSEQR mutation positive samples [the left-hand portion of the table shows known sample information and the right-hand portion (coloured) represents the mutation call output generated by Mutation Surveyor].

4.3.5.3. Screen snapshots of true bi-directional false negatives:

- i) The bi-directional false negative for this data set was c.115-2A>C for sample S19. The variant is picked up on one set for sequence fragments (fragments 4 F/R - variant highlighted in blue) but is totally missed in a second set (fragments 6 F/R) that cover the same portion of sequence. As noted above this data set is generated with multiple primer pairs, where some of the sequences with different primers may overlap, which is the case for this mutation.

Mutation Surveyor is said to analyse/output the data from such a data set best in the whole gene output table where these overlapping regions can serve as internal controls and when overlaps occur, the mutation detection in the overlapping region should be very accurate, that is not the case here as the mutation is only picked up in the data presented with one primer pair and missed in the other.

Given that this mutation is a mosaic, it could be that it falls short of detection thresholds in the traces from fragments 6 F/R. SoftGenetics recognise this sample as an example of a mosaic with a low dropping factor in both directions and that it would be detected/highlighted if one selects the 'Check 2D small peaks' option from the display settings. Following this recommendation this sample was re-analysed under this 'Check 2D small peaks' setting and the mutation was then highlighted in the graphical view with the presence of a green square above the mutation peak in the mutation electropherogram:

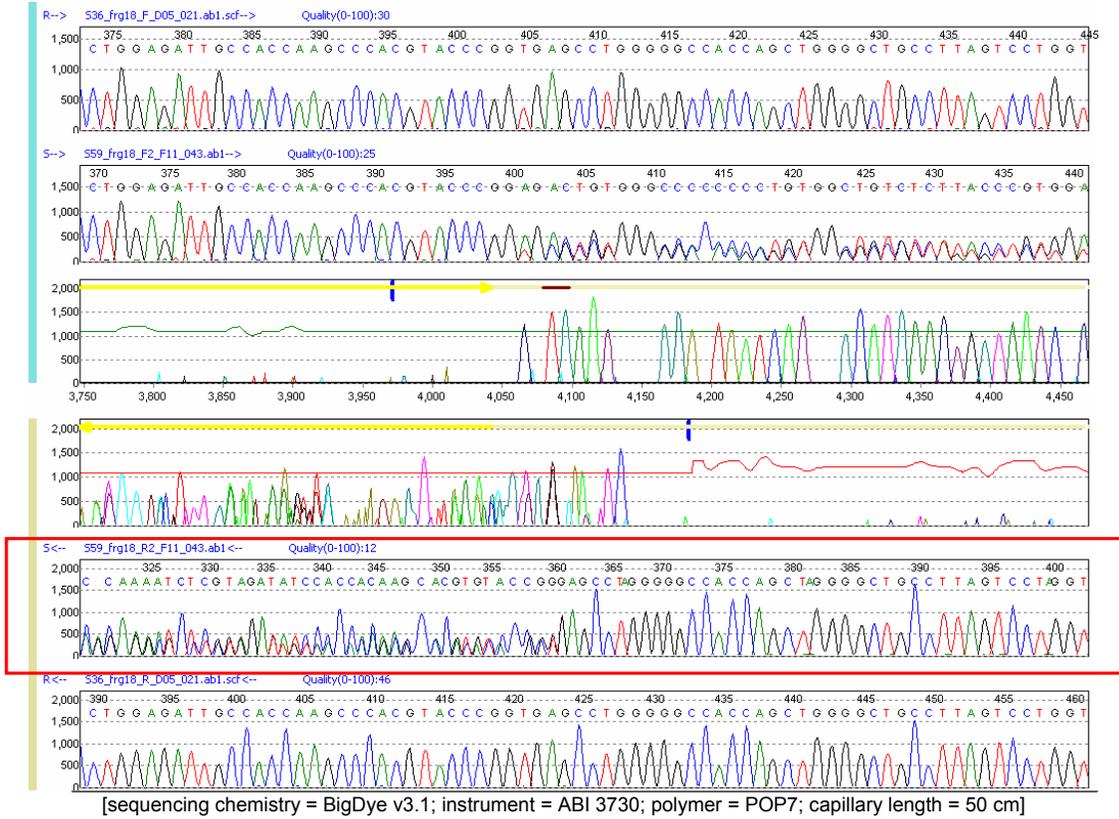


[sequencing chemistry = BigDye v3.1; instrument = ABI 3730; polymer = POP7; capillary length = 50 cm]

Note that this re-analysed mutation is only highlighted in the graphical view and hence the user is still required to manually check the data to confirm these highlighted mutations as genuine mosaics. Furthermore if the user was to rely on this non-automated mutation detection to mine the output for potential mosaic mutations this would greatly increase the number of false positives the user would have to check as small amounts of noise in either the F/R reads of the test sample are frequently highlighted as potential mutations.

4.3.5.4. Screen snapshots of true uni-directional false negatives:

- i) One of the uni-directional false negatives (boxed in red, S59_frg18_R, mutation IVS1446+4_1446+5het_delAG) was a frameshift mutation that was only detected in the forward orientation and missed in the reverse:



SoftGenetics have commented on this uni-directional failure and have noted that this type of mutation has much improved detection in the recent version of Mutation Surveyor (v2.61)

4.4. NF2 Exon linked sequencing data set

4.4.1. Data Source

This is a large data source compiled from our in-house diagnostic mutation scanning service for the NF2 gene using exon linked sequencing (Wallace *et al.* 2004). Meta-PCR is a technique that links multiple exon sequences together into a single chimeric DNA molecule, which is then directly sequenced by a long read sequencing protocol. The data used represents a sample of real patient data produced during three years of service provision (2001-2004).

The data includes many highly variable mutation:normal allele ratios in samples with mosaic mutations and varied types of mutations including large frameshifts which provides challenges for mutation detection software.

Thirteen panels of exon linked sequence data was subjected to bi-directional sequencing. Each analysis panel comprised around twenty two patient samples and a few confirmation/repeats from previous panels or ongoing work, this equates to approximately 25 patient samples per panel or 325 patients in total for this data set.

4.4.1.1. Composition and coverage of the NF2 Exon linked data

Exon linked fragment (size)	Exons covered	Direction	Mutation Surveyor output					
			Ave. size	Ave. Fragment Quality	Number of Patients	TOTAL Length of Sequence Data		
1 (928bp)	1-4	F	889	52	1	7.35kb		
		R	889	46				
2 (1063bp)	6-10	F	985	45				
		R	1010	41				
3 (973bp)	11-14	F	930	45				
		R	929	47				
4 (905bp)	15-17 + 5	F	860	49				
		R	865	49				
Whole data set	-	-	-	-			325	2.41Mb
Whole data set ROI	-	-	-	-				1.51Mb

Table 13. NF2 Exon linked fragments and an estimate of sequence length for the whole data set (Ave. size = an average of the sequence size [defined as size in the Mutation Surveyor output table], calculated from 10 samples; ROI = coding sequence + 15bp either side of intronic sequence).

4.4.1.2. Instrumentation and sequencing chemistry

Panels 1 and 2 were sequenced using an ABI377 slab-based sequencer and BigDye v2.0 chemistry. Panels 3 to 9 were also sequenced using BigDye v2.0 but were analysed on an ABI3100 capillary-based instrument with an 80cm capillary array. From panel 10 onwards the samples were sequenced on an ABI3730 capillary sequencer using BigDye v1.1 sequencing chemistry on a 50cm capillary array.

4.4.3. Analysis settings

Default settings were used as outlined in section 3.9 - figure 14, with the option selected to only display those mutations detected within exons and 15bp of flanking intronic sequence. The 'BasePatch' option was also selected, as this was advised by the technical support at Soft Genetics for long-read sequence data. 'BasePatch' corrects for base calling errors caused by poor mobility correction (this option allows detection of mutations where the mutation threshold for score is unmet due to mobility shift, but overlap and dropping factor are sufficient).

4.4.4. Timing and work load

Each panel of around 25 test samples (13 panels in total) was prepared for automated analysis including a set of wild type sequence trace files and a NF2 GenBank file (derived from a constructed exon linked contig sequence - see section 5.1. *Contig Alignment, Reference Sequences and Automated Mutation Naming in Mutation Surveyor*). As with the previous data sets each batch generated its own text output file of the detected mutations along with a analysis project file.

- Timing of automated analysis for data from a single panel of samples was approximately 6 minutes.
- The output mutation table was then checked against the project file to confirm the identified mutations, this manual check took around 15 minutes depending on the number of mutations reported by the software.
- Total for 13 panels (1.51Mb - based on ROI) = 3.25 man hours OR
- Average analysis time per 10Kb of sequence data = 1.3 minutes

4.4.5. Mutation Detection

Sequencing Platform		Panel	Mutations							
Instrument	ABI BigDye Chemistry		Detected	Expected per strand sequenced	Correct	False Negatives	False Positives	Explainable False Positives	True False Positives	
ABI377	v2.0	1	38	14	14	0	24	6	18	
		2	28	16	13	3	15	7	8	
3		34	20	17	3	17	9	8		
4		12	8	6	2	6	4	2		
ABI3100		5	55	28	27	1	28	9	19	
		8	77	12	10	2	67	31	36	
		9	47	18	17	1	30	23	7	
ABI3730		v1.1	10	59	14	12	2	47	36	11
			11	48	14	12	2	36	29	7
	13		51	24	16	8	35	22	13	
	15		14	10	9	1	5	4	1	
	17		71	10	10	0	61	48	13	
	18		46	16	10	6	36	21	15	
		TOTAL	580	204	173	31 (15%)	407 (199%)	249 (122%)	158 (77%)	

Table 14. Summarised results of expected and observed mutation detection for each panel [% values in brackets = false negative and positive rates expressed as a percentage of the expected number of mutations in a single sequence orientation]

Mutation Type	Number
Substitution	122
Deletion	44
Insertion	4
Insertion and Deletion	3
TOTAL	173

Table 15. Summary of mutations correctly detected

Upon visual inspection of the detected mutations, a number of the false positive results can be explained by poor data (total 249 in table 14), 38 of these were observed to be due to poor quality sequence data, whereas the remaining 211 were due to a sequencing artefact.

N.B the Mutation Surveyor Quality values for the majority of the false negative samples, both those classed as bi-directional (table 17) and uni-directional (table 18), are within the normal range and indicate good quality data (average = 37).

4.4.5.1. False negatives broken down by directional coverage

Bi/uni-directional	Explanation on Visual Inspection	Number of Negatives
Bi-directional false negatives	<i>low level mosaics</i>	5
	<i>true bi-directional false negative</i>	1 (1%)
Total		6
Uni-directional false negatives	<i>misaligned data</i>	3
	<i>mutation masked by frameshift in cis</i>	1
	<i>loss of resolution</i>	1
	<i>true uni-directional false negative</i>	14 (6.9%)
Total		19

Table 16. Breakdown of false negative results [% values in brackets = true false negative rates expressed as a percentage of the total expected number of mutations per strand sequenced for this data set]

4.4.5.1.1. Bi-directional false negatives

Panel	Sample ID	Expected Mutation	Type	State	Ave. Mutation Surveyor Quality	Sequence Fragment [5'>3']	Comments	Fig 20
4	014953	c.1228C>T, p.Q410X	sub	mosaic	57		the mutation electropherogram shows a spike indicating the presence of a mutation and the mutation peak in one of the directions is clearly visible, however these just fall short of the level required for automated detection.	1
11	030780	c.1375C>T, p.Q459X	sub	mosaic	30			2
13	025007	c.448-1 G>A	sub	mosaic	41			3
13	032509	399C>A, p.C133X	sub	mosaic	35		low level mosaics showing very low level peaks in the mutation electropherogram	4
18	922275	343C>T, p.Q115X	sub	mosaic	29			5
18	043593	934_988del55bp	del	mosaic	21		a portion of this mosaic frameshift mutation can be clearly seen in both directions and with clear mutation electropherogram spikes	6

Table 17. Summary of bi-directional false negative mutations [also see figure 20 below for the sequence data and mutation electropherograms] the sequence fragment idiograms illustrate the position of the missed mutation on the fragment; comments in red denote true false negatives - with no explainable observations; the exact percentage mosaicism has not been quantified for these mutations}

Like the bi-directional false negative mutation in section 4.3.5.3, these mosaic mutations have been missed by automated mutation detection, but are highlighted when one selects the 'Check 2D small peaks' option from the display settings (see section 4.3.5.3. for further comments)

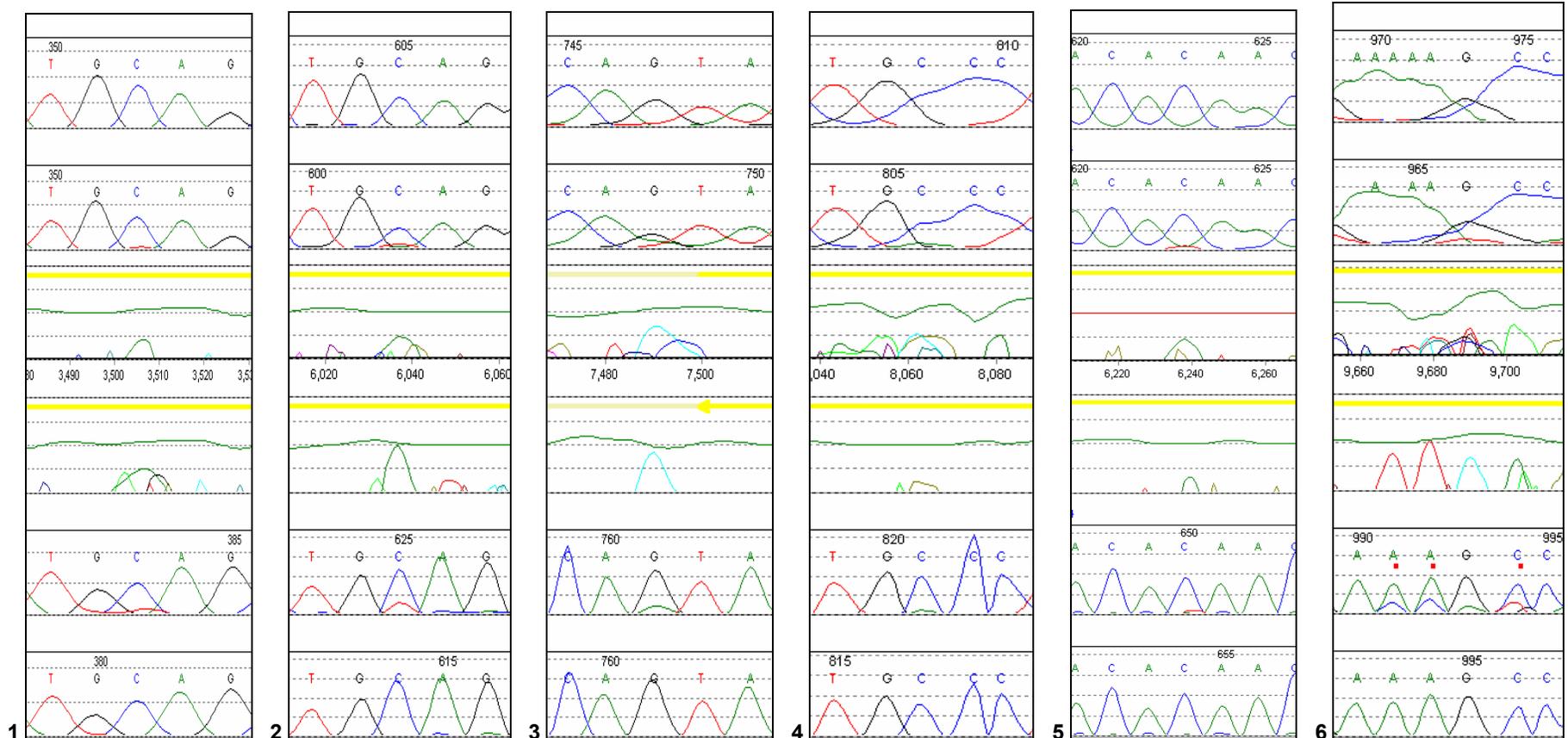


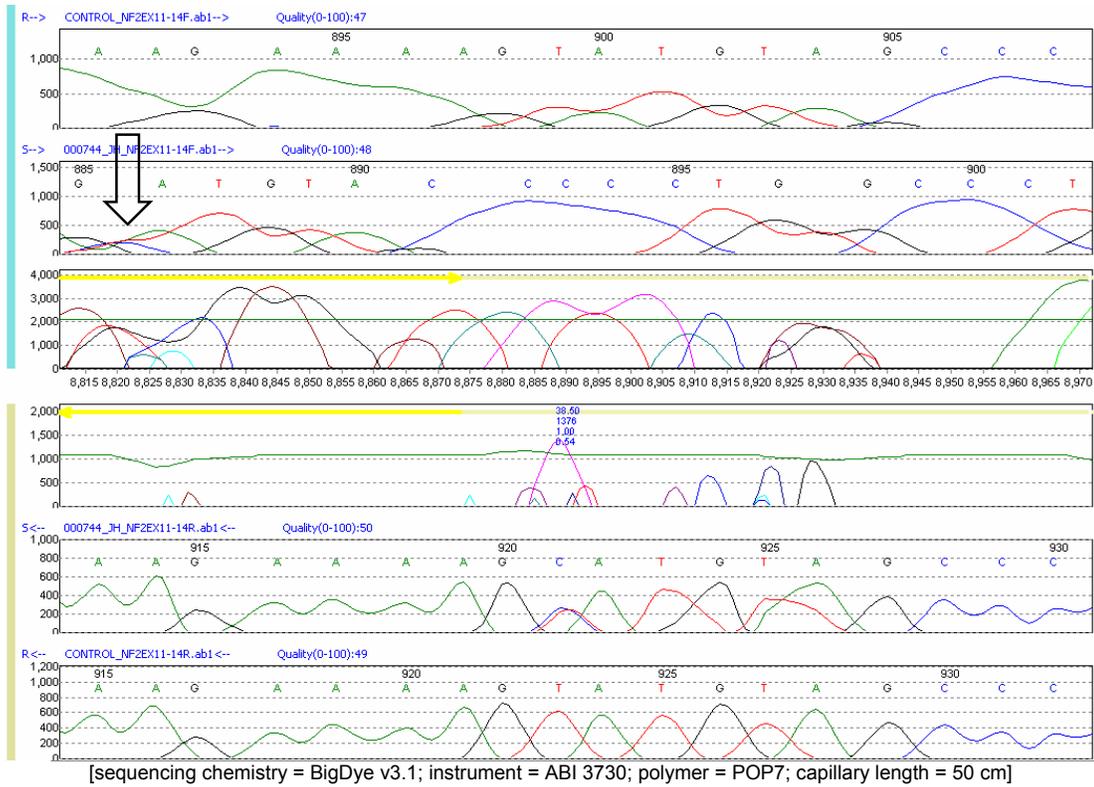
Figure 20. Screenshots from Mutation Surveyor of the bi-directional false negative mutations [top and bottom plane = F and R of wild type control trace; one in from top and bottom plane = F and R of test sample trace; middle plane = the mutation electropherogram].

4.4.5.1.2. Uni-directional false negatives

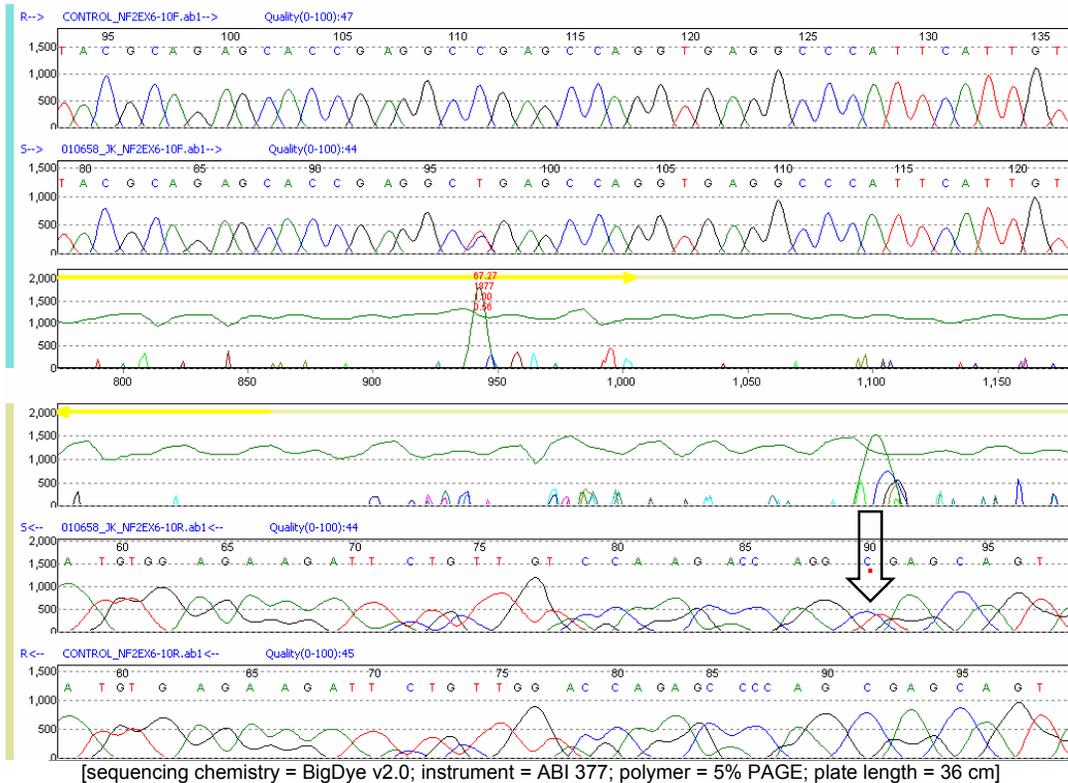
Panel	Sample ID	Expected Mutation	State	Dir.	Mutation Surveyor Fragment Quality	Sequence Fragment [5'>3']	Comments
2	012414	c.144_154del11_ins7	heterozygous	R	39		all insertion/deletion mutations located towards the end of the sequence read
3	014009	c.57_58insA	heterozygous	R	50		
3	992998	c.599_600insG	heterozygous	R	50		
8	024011	c.441delG	mosaic	F	52		
9	025047	c.288_290delCCT	heterozygous	F	36		
10	030071	c.1584delC	heterozygous	R	33		
10	030564	c.448-20del25	mosaic	F	32		
13	030701	c.1571_1572delAA	mosaic	F	38		
13	032342	c.672delC	heterozygous	R	31		
15	040072	c.1503delT	heterozygous	F	34		
18	040431	c.396_412del17bp	mosaic	F	31		
2	000744	c.1574+2T>C	heterozygous	F	48		mutation contained within a misaligned terminal portion of sequence fragment respective wild type trace
2	010658	c.592C>T, p.R198X	heterozygous	R	44		mutation contained within a misaligned portion of sequence fragment to the reference sequence.
3	990004	c.810G>A	mosaic	F	68		mutation lies 3' of a frameshift due to a second mutation (c.737delC)
5	020738	c.447+2T>C	heterozygous	F	61		no observable problem a true false negative – although towards the end of a read
8	023940	c.592C>T, p.R198X	heterozygous	R	40		mutation contained within a misaligned end portion of sequence fragment to respective wild type trace
13	025007	c.958C>T, p.Q320X	mosaic	F	31		a slight loss of resolution but visible by eye
13	032327	c.447+1G>C	heterozygous	R	37		no major observable problem but a slight loss of resolution
18	042898	c.1340G>A, p.R447K	heterozygous	F	25		no observable problem a true false negative

Table 18. Summary and comments on uni-directional false negative mutations [also see screen snapshots below; sequence fragments illustrate the position of the missed mutation on the fragment ; comments in red denote true false negatives - with no explainable observations; the exact percentage mosaicism has not been quantified for these mutations }.

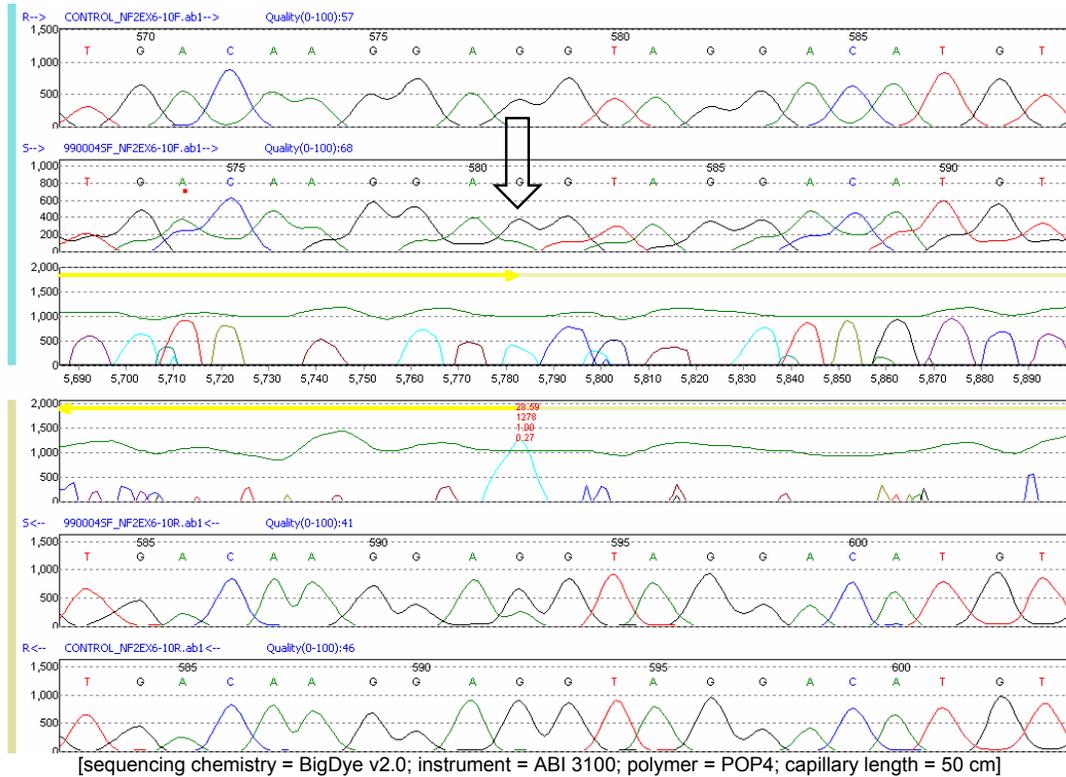
- i) Sample 000744 (c.1574+2T>C): the mutation (indicated by the black arrow) appears to be missed in the forward orientation as it is incorrectly aligned to its respective wild type normal control trace. This misalignment was seen to be present for around the last 100 bases of the sequence. SoftGenetics say that this is caused by bad basecalling and poor quality in the forward trace:



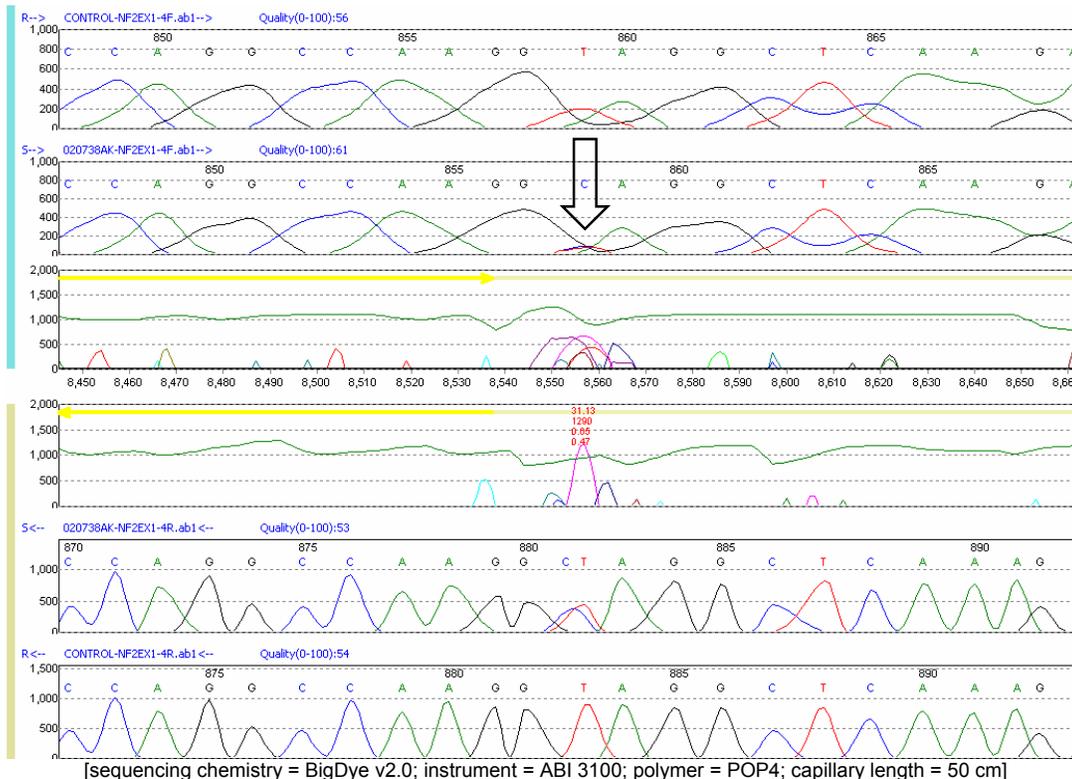
- ii) Sample 010658 (c.592C>T, p.R198X; see section 4.1.5.3. ii, for a comment on the red dot seen above the missed mutation): again this electropherogram exhibits a false negative in a misaligned segment of sequence, where the sample along with its respective normal control sequence are both misaligned relative to the reference sequence. The peak from the mutant allele is clearly present. SoftGenetics comment on this as poor quality sequence in the reverse and that Mutation Surveyor v2.61 will eliminate the poor quality data for mutation detection.



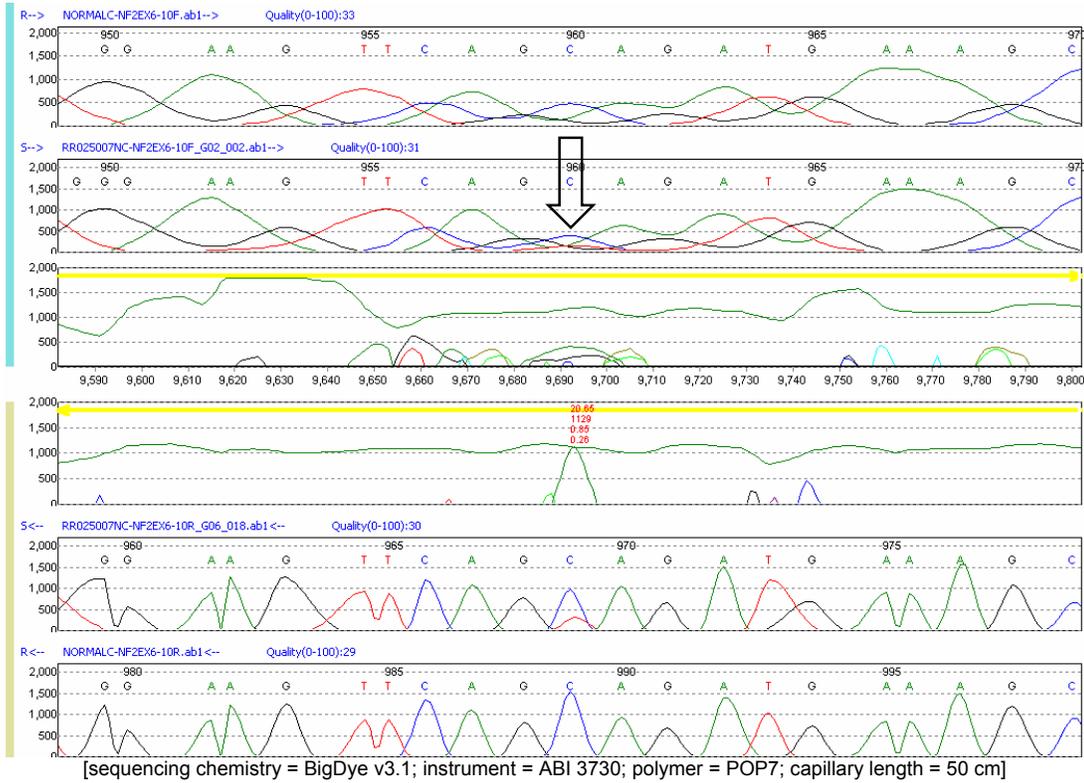
- iii) Sample 990004 (c.810G>A): this mutation is masked by a shift (caused by the second compound heterozygote mutation, c.737delC), however the shift in the sequence of one base pair does not obscure the mutation completely which is still visible (arrowed):



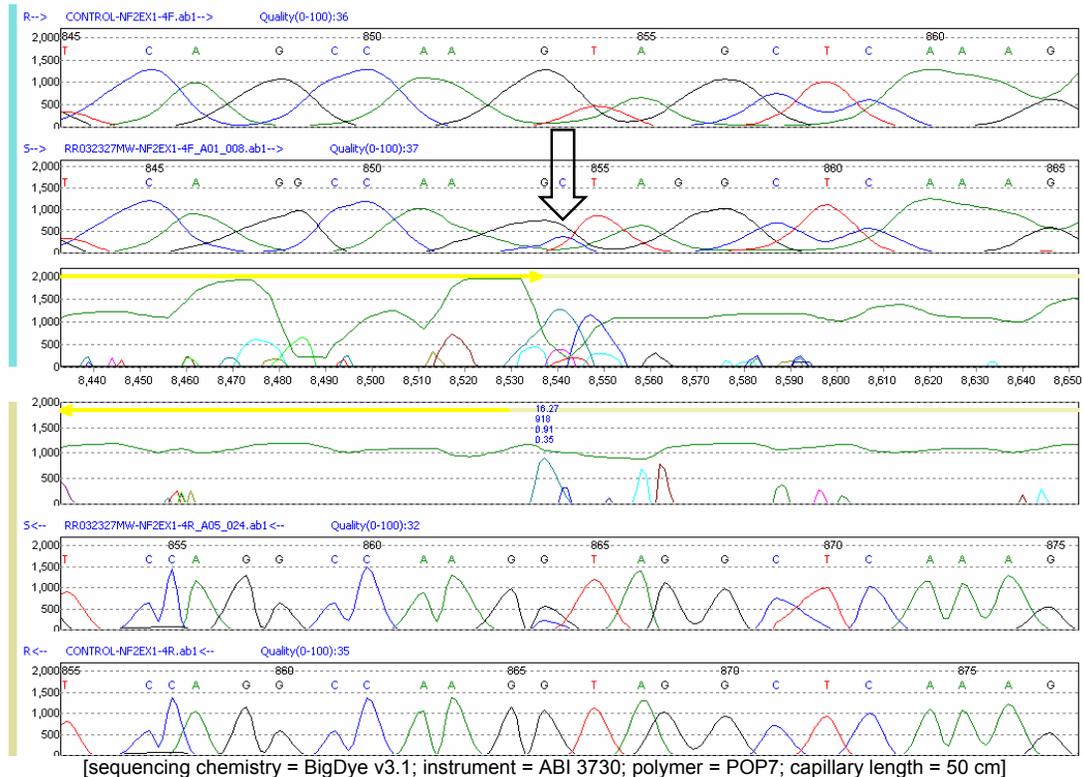
- iv) Sample 020738 (c.447+2T>C): a true false negative (no observable problem with the sequence data), the mutant base has been called as a C (arrowed) but it has been missed by automated mutation analysis as a true mutation in the forward orientation:



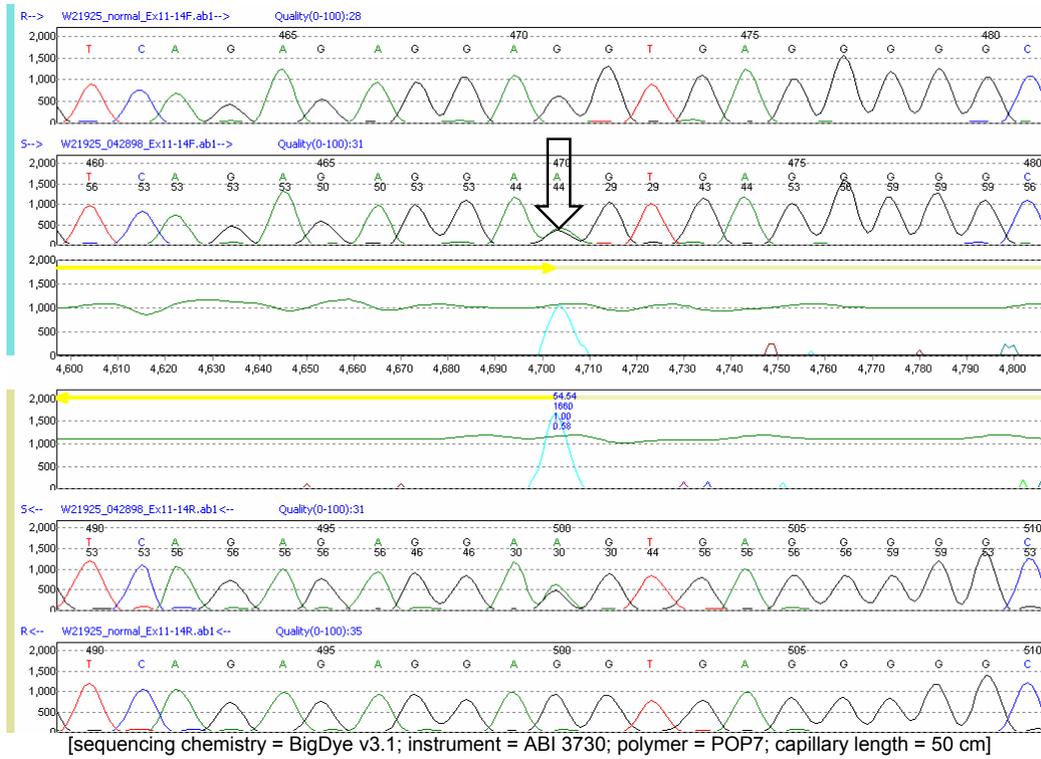
- v) Sample 025007 (c.958C>T, p.Q320X): mutation missed in the forward strand. There seems to be no major observable problem with the sequence data except a slight loss of resolution. SoftGenetics say that this is probably due to poor data, however this is not marked as being poor data and has been accepted by the software under default conditions:



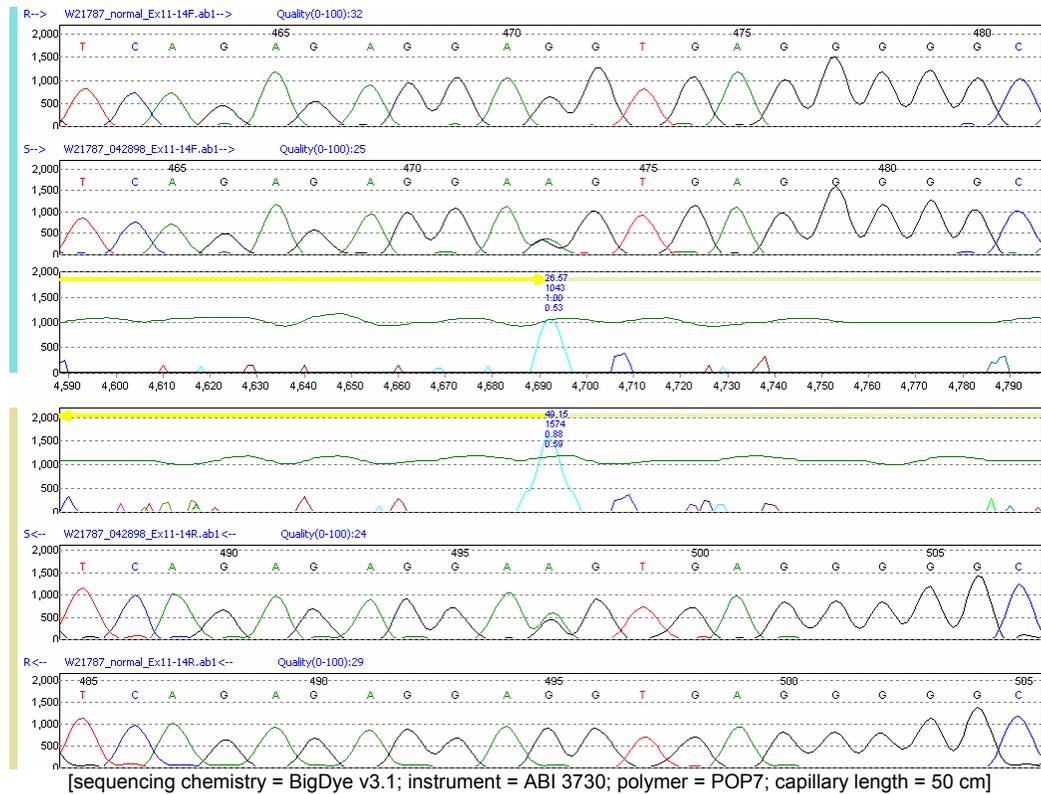
- vi) Sample 032327 (c.447+1G>C): again no major observable problem but a slight loss of resolution. Once again SoftGenetics say that this is probably due to poor data, however this data has been accepted by the software for analysis:



vii) Sample 042898 (c.1340G>A, p.R447K): the mutation is clearly observable in both orientations (with equal height) this is an unexplained false negative:



This re-sequenced repeat sample of specimen 042898 (below) shows the same indistinguishable sequence trace result as the original sample analysis, and the automated mutation detection by Mutation Surveyor has now noted the mutation in the forward orientation. This is a good example where the results of Mutation Surveyor analysis can be divergent for a sequencing repeat even though the quality scores of both sets of sequence data can be very similar:



SoftGenetics acknowledge that this mutation at the top of the page was missed due to a problem in the software's coding, and that they will be checking this further and improve on this in future versions.

5. OTHER FEATURES OF MUTATION SURVEYOR

5.1. Contig Alignment, Reference Sequences and Automated Mutation Naming in Mutation Surveyor

The ability to automate the mutation naming process in Mutation Surveyor is an extremely useful facility as manual naming of mutations is a time consuming and error prone process. Consequently a fully functioning reference sequence is important for the deployment of Mutation Surveyor as this dictates the accuracy of downstream processes such as the naming of identified sequence changes.

Mutation Surveyor has a feature for downloading GenBank files from a database compiled and held by SoftGenetics for sequence analysis (*N.B.* users should note that GenBank make no claims as to the accuracy of the data in their files). When the operator inputs sample files to be analysed and leaves the GenBank file and reference trace inputs blank, the software automatically queries the Mutation Surveyor database comprised of GenBank sequences taken from the NCBI website, and matches the test sequences with the most recent GenBank file within the SoftGenetics database.

We attempted to use this feature of Mutation Surveyor on a number of occasions for each of the data sets in the study. However Mutation Surveyor failed to return a GenBank file. SoftGenetics point out that this type of difficulty will occur when the user has poor quality traces for comparison to the GenBank file and the way to combat this is to have several high quality traces for comparison to the GenBank file. In addition in v2.61, the software no longer sends the poor quality traces to retrieve a GenBank file from the SoftGenetics website.

Given the failure of Mutation Surveyor to download an appropriate GenBank reference file, sequences were manually downloaded from the NCBI web site. Text or GenBank format reference sequence files were then opened in Mutation Surveyor with the built in GenBank file editor tool and checked to see that the coding sequences and the exon/intron boundaries were correctly mapped and that the correct protein translation was made.

Most of the GenBank files used in this study gave good and consistent results with the exception of the Exon linked sequencing data-set. For the exon linked and the VariantSEQr NF2 data-sets a number of different types of reference sequence were tried to establish the most reliable type of reference input format for correct sequence alignment and numbering of these two comparable but different types of data (table 19).

Two groups of files were tested, GenBank files and single exon files. Single exon files are generated within Mutation Surveyor, where GenBank files are opened and then saved as single exon files (the larger GenBank sequence is broken up into exons plus 300bp of sequence either side) these can then be viewed and checked using the sequence file editor within Mutation Surveyor. Such single exon files could be useful for analysing a mixed population of samples for a number of different sequenced fragments.

5.1.1. VariantSEQr data set

On testing with the VariantSEQr data set, only one of the reference files carried out mutation naming acceptably. This reference file type 1 (table 19 - GenBank file based on NCBI35:22:28324117:28419) gave accurate and consistent results. As expected this GenBank file of a genomic NF2 sequence annotated with cDNA information should have encountered no problems with sequence fragments covering the NF2 gene region. However the single exon version of the same reference file (reference file type 4) encountered problems with cDNA numbering.

Reference file type 2 is a GenBank file based on a construct of full exons plus 100bp of intronic flanking sequence, however the VariantSEQr data did not always align correctly, as some of the sequence fragments contain >100bp of intronic sequence and Mutation Surveyor tended to force such fragments to align with the available reference sequence (this problem could easily give rise to false positive calls as illustrated in the screen snapshots below).

The reference files derived from the exon linked sequence construct were incompatible with the VariantSEQr data. This was because they include foreign sequence such as primers/linkers from the exon linked process, which cause misalignment of sequences and leads to the calling of false positive mutations within the defined ROI.

For example, for the VariantSEQr data using reference file type 2 a "true" mutation has been identified within 15 bp of the intron/exon boundary. Although this was a mutation outside the ROI the software forced the sequenced fragments to align with a non-homologous portion of the reference sequence leading to the mutation being misclassified as within the ROI (figure 21).

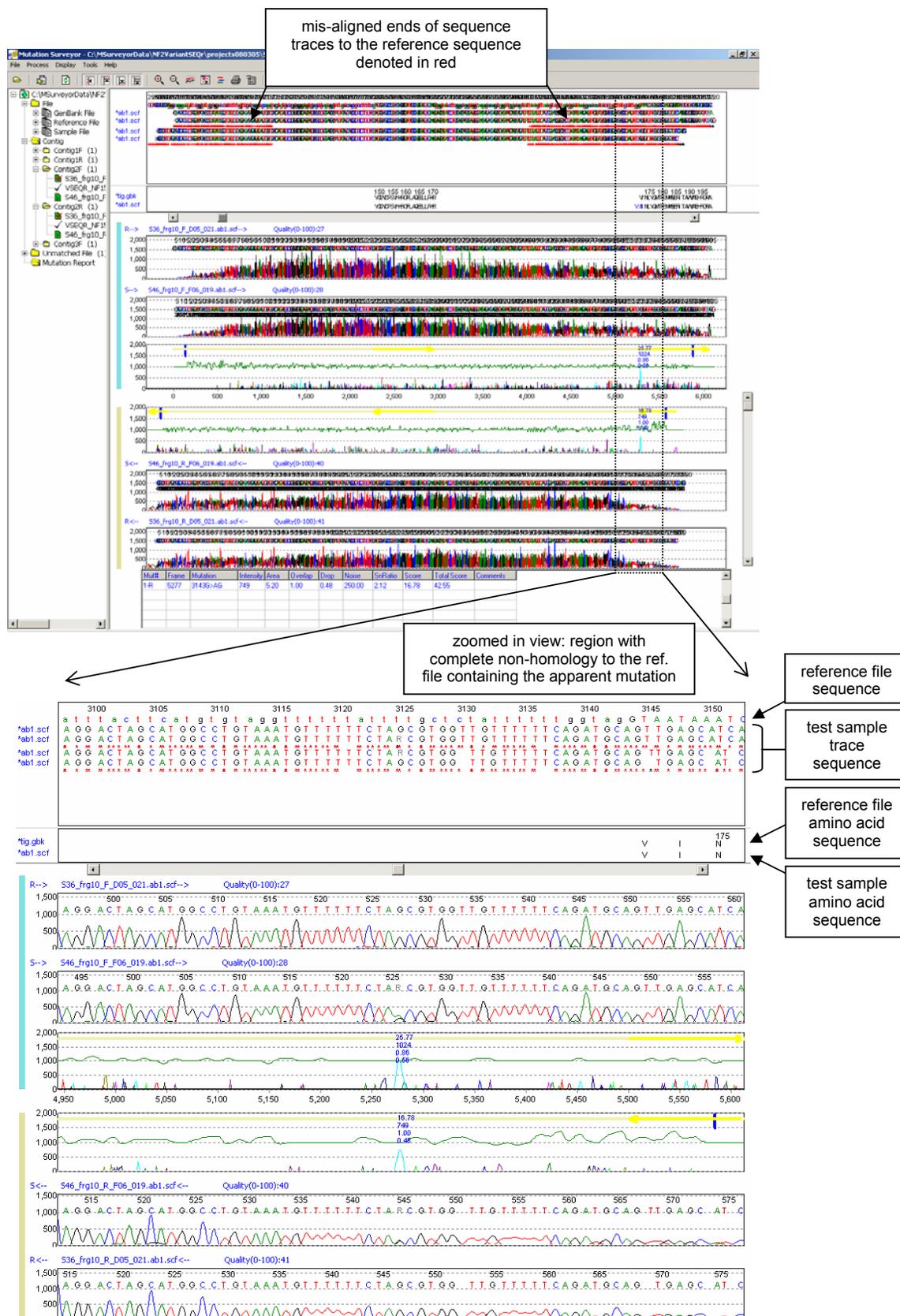


Figure 21. Screen snapshots from Mutation Surveyor of misaligned trace data and a false positive call within the ROI [the upper screen snapshot illustrates the mis-alignment of the trace peripheral sequences; the lower screen snapshot shows the mutation electropherogram for this apparent mutation that has been detected within 15bp of the intron/exon boundary, due to the forced alignment to the reference file sequence].

Reference Type	Data-set & Feature						
	NF2 VariantSEQr			NF2 Exon linked			
	Amino Acid / cDNA Numbering	F / R Sequence Fragment Pairing	Sequence Alignment / Text View	Amino Acid / cDNA Numbering	F / R Sequence Fragment Pairing	Sequence Alignment / Text View	
GenBank File	(1) GenBank file based on NCBI35:22:28324117:28419	✓	✓	✓	✗	✓	✗
	(2) GenBank file based on a construct of full exons + ~100bp intronic sequence either side of these	✓	✓	✗/✓	✗	✗	✗
	(3) GenBank file based on exon linked contig of full exons + ~20bp intronic sequences either side of these, inc. primer / linker sequences	✗	✓	✗	✗	✓	✓
Single Exon Files	(4) Exons extracted by Mutation Surveyor from file (1) above	✓/✗	✓	✓	✓/✗	✓	✗
	(5) Exons extracted by Mutation Surveyor from file (3) above	✗	✓	✗	✗	✓	✓/✗
	(6) Exon files with only 20bp of intronic sequences either side	✓/✗	✓	✓	✓/✗	✓	✓/✗

Table 19. Tabulated results of tested reference files with VariantSEQr and exon linked sequence data.

5.1.2. Exon linked data set

No single type of reference sequence template fulfilled all the desirable output characteristics for the exon linked data set. The problems encountered by Mutation Surveyor in trying to use any one of the outlined reference sequence types (table 19) for this data set included failure to pair F/R traces, misalignment of traces/incorrect text output view (see section 3.5. *User Interface - figure 8*) and incorrect numbering for amino acids and cDNA.

On a few occasions samples did not pair up for bi-directional analysis (this was most pronounced for the largest fragment encompassing exons 6-10). SoftGenetics technical support recommended selecting the “BasePatch” option within the contig settings window to minimise the problem of mis-pairing samples and any misalignment of samples to the reference sequence. The “BasePatch” option corrects for base calling errors caused by poor mobility shift (this option allows detection of mutations where the mutation threshold for score is unmet due to mobility shift but overlap and dropping factor are sufficient). This did improve the alignment of sequences and their pairing, but a few samples still persistently mis-paired. The length of the fragments which proved difficult to pair was at the extreme of those likely to be analysed by a typical Mutation Surveyor user but our experience does illustrate the limits of the contig assembly algorithm.

Mutation Surveyor encountered great difficulty in dealing with this data and producing correctly aligned data and a text output view (see section 3.5. *User Interface - figure 8*). For reference types 2 and 4 the Mutation Surveyor application freezes and then stops responding when trying to view the text-output view. In summary the Mutation Surveyor text-output view will not handle sequence data against reference sequence templates that are not fully compatible with the sequenced data. Also none of the templates are able to correctly number the sequence data by cDNA order, as the exons are not in consecutive order.

In conclusion the exon linked derived GenBank reference file (type 3) was the most suitable type of reference sequence to get correct pairing and alignment of sample data at a cost of not getting any correct numbering information (cDNA and protein). We do accept however that exon linked data is a specialised form of sequence data and is only relevant to a small minority of Mutation Surveyor users.

5.2. Strengths of Mutation Surveyor:

- Mutation Surveyor is a comprehensive sequence analysis program incorporating many features that are useful to a diagnostic laboratory
- Mutation Surveyor reads GenBank files (in a .txt or .gbk file format) and converts them simply into a usable reference sequence backbone. The software also has a GenBank file editor, allowing the user to add further features such as ROI information and position of polymorphisms or frequent mutations.
- Allows the conversion of GenBank files into single exon reference sequence files which can then be edited in the sequence editor allowing the user to change the amino acid numbering (start point) for any given ROI CDS. This feature may be useful for laboratories analysing sequence data generated as single exon sequence fragments.
- A number of different wild type control sequences/traces can be loaded up into the software, and then Mutation Surveyor uses only those traces with the best quality scores.
- Mutation Surveyor automated mutation detection algorithm has a high sensitivity, especially for frameshift mutations. A very useful feature of the software is the de-convolution of frameshift heterozygotes into separate alleles thus reducing the need for time consuming manual analysis.
- Given an accurate GenBank file as a reference backbone, the majority of mutation calls are either named correctly or deviate by a few bases. However, with the naming of frameshift mutations, names given to mutations on the forward orientation tend to be correct whereas those given to the mutations in the reverse orientation tend to be consistently incorrect.
- There is a file renaming tool which allows the user to edit file names in bulk facilitating better pairing of F/R direction files.
- Rapid processing of data; batches of 6 samples of the NF2 VariantSEQR data (equating to ~191kb of sequence data only took 8 minutes to process.
- Within the graphical interface of the software, a selection of sequence electropherograms can be easily highlighted using the mouse and then zoomed in. This feature is extremely useful when navigating through sequence data.

5.3. Limitations of Mutation Surveyor:

- Automated mutation detection by Mutation Surveyor exhibits reduced sensitivity for low level mutations, such as mosaics. Therefore data potentially containing mosaic mutations should be visually checked (aided by the 'Check 2D small peaks setting') to reliably detect minority alleles (Mutation Surveyor is quoted to have a >99% accuracy, with a sensitivity down to 5% of the primary peak - when sequence data meets a minimum Phred score of 20. Since Mutation Surveyor does not provide Phred quality scores it is unclear what proportion of our data sets meet this requirement).
- In uni-directional mode a detection rate of $\geq 95\%$ with a mutant allele sensitivity of 10% of the normal allele is claimed. Although data solely comprising uni-directional reads has not been assessed in this study (under uni-directional settings/thresholds), a lower sensitivity for uni-directional data is evident from the true uni-directional false negative rate observed for the data in this study (ranging from 0.2-6.9% - table 21). Based on these preliminary findings, we would not recommend analysis reliant solely on automated mutation detection in single orientation reads, but that assessment of a set of high quality uni-directional sequences (tested under uni-directional thresholds) would be a useful follow-up to this study.
- Data produced on the Beckman Coulter CEQ8000 platform using the CEQ DTCS chemistry is analysed less effectively by Mutation Surveyor. False positive and negative rates are increased. SoftGenetics recognise this as a problem due to the sequencing chemistry and in order to solve this problem they will need to work with the manufacturer to overcome this limitation.
- SoftGenetics accept that Data from MegaBace systems will often result in false negatives and false positives with indel detection, which they claim is due to the way MegaBace artificially alters the DNA migration time.
- On several occasions when a single patient specimen has been sequenced more than once, and all the samples have been analysed, the results of Mutation Surveyor analysis can be divergent even though the quality scores of both sets of sequence data can be very similar. In response to this SoftGenetics stress the importance of maintaining consistent conditions/settings and controls for every batch of data analysed using Mutation Surveyor. We feel that these requirements have been met and that consistency between different batches of data has been maintained. One possible explanation for such divergent results could be that there is some variation between different batches that result in data being slightly below the detection thresholds of Mutation Surveyor.

- Although the de-convolution of frameshift mutations is a very useful feature, this is often only successfully achieved in one sequence orientation. SoftGenetics say that this feature has been greatly improved in the 2.61 version of the software.
- Mutation Surveyor is unable to automatically identify mutations reliably downstream of a de-convoluted frameshift. A failure to do this was noted on several occasions. Mutation Surveyor consistently failed to do this for the mutation c.1408AG>A which is upstream of a frameshift mutation c.1519_1521delATC, in the EMQN data set. It also failed to detect a c.810G>A substitution downstream of a c.737delC mutation. This particular patient specimen was present in both the VariantSeqr and exon linked data sets and in both cases c.810G>A was not detected through the de-convoluted trace of Mutation Surveyor. SoftGenetics say that this feature is difficult to perfect due to the challenge of automatically de-convoluting this type of mutation and that they are much more accurately called in v2.61 of Mutation Surveyor and they are working to enhance this feature in future versions.
- It is difficult to determine when a portion of the ROI has not been screened by the mutation detection algorithm, as there is no automatic monitoring of successful ROI coverage. The limits of analysis by the mutation detection algorithm are indicated on the mutation electropherogram by small vertical blue bars and tabulated in the output table (figures 22 and 23).
- The software has a very low threshold for data of acceptable quality. Using the default settings (quality threshold =0) very rarely does Mutation Surveyor give poor data a score <0 as most data meets the threshold of 0. Whereas 'Bad Data' is clearly highlighted in the mutation column and is signified by a -1 value in the mutation number column. The distinction between 'Bad Data' and 'Low Quality' is that the 'Bad Data' is not assembled into a contig as it is designated 'unmatched'. Consequently Mutation Surveyor accepts data of a wide quality range and only rejects those that are very poor quality or unmatched to the reference sequence. Furthermore Mutation Surveyor has lower sensitivity in low quality data although data of this quality meets the Mutation Surveyor default settings and is accepted by the software. SoftGenetics responded to this as follows: "In most instances, the defaults set in the program will yield the best analysis results for most users. However, it is possible for the user to set the lane quality threshold to whatever value they desire. However, by raising this value much above 10, the user may be eliminating high accuracy mutation calls in lower quality lanes. The default settings allow for low quality data to be analyzed, although with lower sensitivity than in regions of high quality sequence traces. This is because our goal for the software is to perform a reliable, automated approach to mutation detection that works for most data the majority of the time. However, it is the user's responsibility to determine if they choose to remove low quality traces by raising the lane quality thresholds".
- Mutation Surveyor demonstrated a lower sensitivity in long read DNA sequence data especially towards the 3' end of the read where it has trouble with data that shows loss of resolution. SoftGenetics note that they have now improved on the 'score trimming' method in v2.61 thereby allowing the user to set a threshold for trimming off terminal sequences. This feature should correct for any mobility shifting in the samples.
- No error/trouble shooting guide. Error messages are hard to understand, i.e. failed autorun projects do not generate any information indicating why they failed. A run error log would be useful. However as SoftGenetics point out there are FAQ's on the technical services page at their website. Also they intend to improve this in the future. Additionally, they offer web-based conferences to clients in order to discuss analysis issues.
- Only two sorts of output files can be generated in the autorun mode, a project file and a text results output table (text, in a tab delimited format), there is no choice of what output files can be automatically generated in the autorun mode. SoftGenetics point out that in v2.61 there is also a 'Velculescu' report format that can be saved. All other reports can be obtained manually from the saved project.
- The results output file is just a tabulated text version of the advanced bi-directional report output. Also none of the other output file types available with the application running have any functionality in the way of producing a mutation snapshot or a hyperlink - they are simply tabulated results. SoftGenetics accept that while none of the output reports displays a snapshot of the mutation, it is possible to print the mutations as they appear in the sample lanes by clicking either the Print or Print Clinical Report buttons. Additionally, it is possible to select any portion of the electropherogram for snapshot copying by capturing the screen (Ctrl + Prt Scr), by selecting the portion in the Mutation Surveyor software (Ctrl + Shift + left mouse click draw box, then right mouse click and select copy). Additionally, it is possible to export the images to a Microsoft Word document by clicking the Word icon in the tool bar.
- There are only a limited number of output measures that are generated and tabulated in the output views with little facility to customise or manipulate the way the data is presented. SoftGenetics recognise that this is a short-coming in their software and will be improving this function in the future.
- The clinical reports that can be printed directly for a patient sample are only available through the graphical display and cannot be saved (only available for direct printing). Further to this the clinical report output/format is unlikely to meet the requirements of most diagnostic labs for a clinical report although it may serve as a useful internal record. Again, SoftGenetics are aware of this and respond by saying they are glad to alter the Clinical Report so that it is more applicable to clinical laboratories. However, this will

take time and input from relevant users in order to implement changes that make sense for the common clinical user.

- The scroll bar control in the top pane in the graphical view is difficult to control, it does not move with the sequence as it is manually scrolled, the bar centres after scrolling and does not stay in position to indicate the location of the current window. Navigating using this control is difficult and the user can lose the sequence data if scrolled too far in either direction of the ROI. Alternatively the user can navigate through the sequence trace data by right-mouse clicking and dragging the plane to the left or right, this movement gives the user better control of the graphical view planes. In fact it is possible to scroll through all graphical displays by right-mouse clicking and dragging to the left or the right.
- Intermittently sequence data apparently correctly aligns to the reference sequence but when the user toggles to text output, the programme freezes/crashes.
- Samples are not always correctly paired. In advanced two directional output this can only be overcome by using a stringent trace file naming convention. Even files that are named according to a strict convention and have a few terminal additional characters indicating capillary number, such as those added by Applied Biosystems DNA sequencers (e.g. filename_B01_007.ab1) are often unpaired during analysis. SoftGenetics point out that there is the file name match editor available from the tools menu and the load 2D match option available to the user where they are having difficulty pairing F/R sequence files.

Figures 22 and 23 illustrate the failure of Mutation Surveyor to give an accurate indication of ROI covered and accepting poor quality data. The tabulated results (figure 22) give no direct indication as to the quality of the analysis for the trace shown in figure 23.

The user is left to interpret the size data, the trace quality score and the graphical trace data for each orientation, before one could reject the Mutation Surveyor analysis and the trace data. A percentage coverage measure in each orientation and a bi-directional coverage indicator would be very useful and give the user confidence that the desired regions of interest have been satisfactorily analysed by the mutation detection algorithm.

No.	Sample File	Gene	Start	End	Size	Quality	Mut#	Mutation1	Mutation2	Mutation3	Mutation4	Mutation5
1	13_MEN1_P1_F.ab1	MEN1	4266	4584	319	23	1	4270_4271insTT				
2	13_MEN1_P2_R.ab1	MEN1	4318	4566	249	41	2	n.a.	n.a.	n.a.	n.a.	n.a.
3	13_MEN1_P2_F.ab1	MEN1	4266	4582	317	37	4		4273T>C,425		4301AC>C,155T	
4	13_MEN1_P3_R.ab1	MEN1	4326	4353	28	0	0	n.a.	n.a.	n.a.	n.a.	n.a.
5	13_MEN1_P3_F.ab1	MEN1	4375	4582	208	5	0	n.a.	n.a.	n.a.	n.a.	n.a.
6	13_MEN1_P4_R.ab1	MEN1	4312	4566	255	40	2	n.a.	n.a.	n.a.	n.a.	n.a.
7	13_MEN1_P4_F.ab1	MEN1	4285	4582	298	10	3	n.a.	n.a.		4300C>AC,155T	4306T>G,157L>
8	MEN1_GenBank_F	MEN1	4303	4563	261	68	1	n.a.	n.a.	n.a.	n.a.	n.a.
9	MEN1_GenBank_R	MEN1	4267	4582	316	87	1	n.a.	n.a.	n.a.	n.a.	n.a.

Figure 22. Two directional output table for an analysis of the MEN1 gene of sample P3 (boxed in red;).

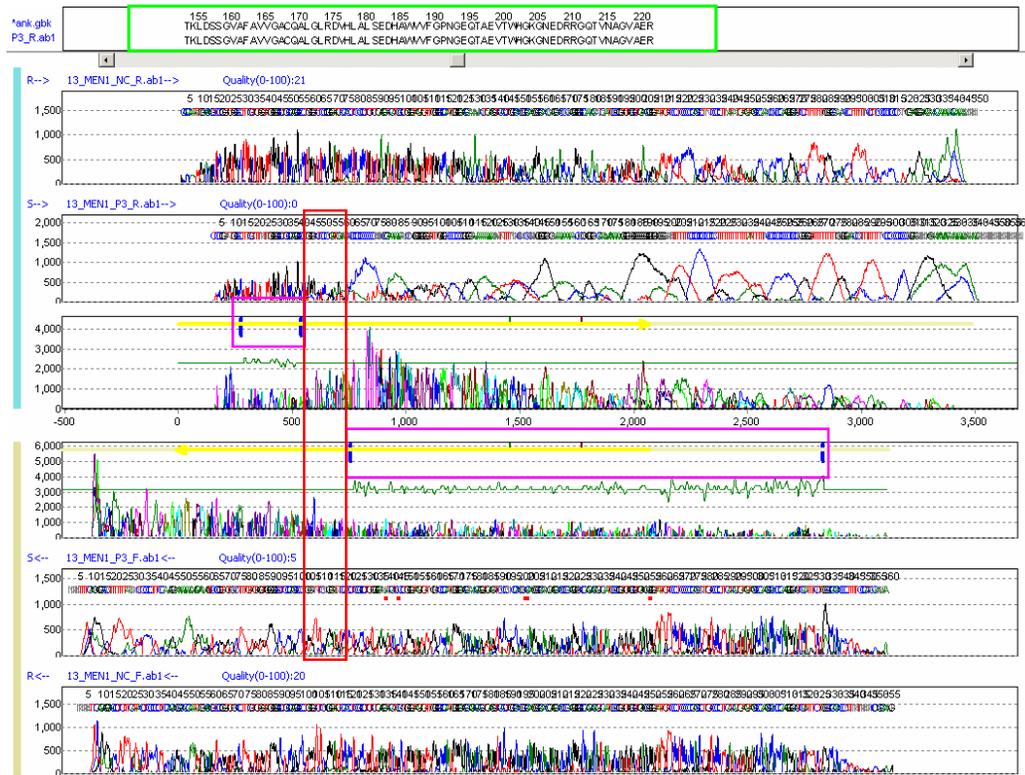


Figure 23. Overview of sample P3 trace data (boxed in green = the amino acid sequence i.e. the ROI; boxed in pink are small vertical blue bars, representing the start and end points of the quality read length/size as given in the tabulated output in figure 22; N.B. how there is no good quality read length to cover both orientations; boxed in red = no coverage in either direction for a portion of the ROI]

6. SUPPLIER DETAILS & CURRENT PRICES

SoftGenetics® DNA sequence data analysis software - Mutation Surveyor™

Available from local distributors or from SoftGenetics where your area is not covered (for a more extensive list of local suppliers including Europe, see SoftGenetics website)

Company/main supplier:

SoftGenetics Inc.
Suite 241
200 Innovation Blvd
State College
PA 16803
USA
phone: 814-237-9340
fax: 814-237-9343
email: info@softgenetics.com
website: www.softgenetics.com

UK supplier:

BioGene Ltd
BioGene House
6 The Business Centre
Harvard Way
Kimbolton
Cambs PE28 0NJ
United Kingdom
phone: +44 (0) 845 1300950
email: info@biogene.com
website: www.biogene.com/index.cfm

Single User Licences						Annual Updates	
Product Name	Comment	Licence Type	Product Code	£	€	£	€
Mutation Surveyor™	400 lane capacity	academic	MS-001	3600	5373	720	1075
		non-academic	MS-001	5852	8734	1694	2528
	400 lane capacity network version	academic	MS-001N	3870	5776	720	1075
		non-academic	MS-001N	6438	9609	1694	2528
	additional user licences for use with MS-001/N	academic	MS-002	2250	3358	450	672
		non-academic	MS-002	3060	4567	563	840
	48 lane capacity	academic	MS-004	2835	4231	900	1343
		non-academic	MS-004	4230	6313	1238	1848
Mutation Explorer™	48 lane capacity with fixed default analysis parameters for diagnostic applications	academic	ME-001	2655	3963	540	806

Table 20. Estimated costing for the various automated Mutation Surveyor™ applications.

7. SUMMARY

To assess the performance of automated mutation detection using SoftGenetics® sequence data analysis software Mutation Surveyor™ v2.51 in a diagnostic setting, we tested four sets of bi-directional sequence data that covered a broad spectrum of sequencing chemistries, laboratories, sequencing platforms and read lengths. We attempted to represent the spectrum of bi-directional sequence data generated in clinical diagnostic laboratories.

In bi-directional mode, Mutation Surveyor is claimed to detect >99% of mutations, with sensitivity to the mutant allele extending down to 5% of the primary peak - when the sequence data meets a minimum Phred score of 20. Since Mutation Surveyor does not provide Phred quality scores it is unclear what proportion of our data meets this standard. However for the data sets in this study after excluding all possible explanations for false negative results through visual inspection of the trace data, the 'true' bi-directional false negative rates depending on the data set ranged from 0.0-4.9% (table 21).

Mutation Surveyor has a lower sensitivity in automatically detecting and reporting mosaic mutations and although detection is claimed to be effective down to 5% this does not appear to be the case on our data where only 62% (33/53) of mosaic variants were detected automatically.

Given these findings even with good quality bi-directional data we would not recommend fully automated analysis be used alone with Mutation Surveyor analysis parameters set to their default values, and that manual review of the data is retained in diagnostic laboratories. However individual laboratories may find that with the fine tuning of the various parameters they can achieve satisfactory results. Data can be manually reviewed within Mutation Surveyor, however this considerably extends the length of time taken to process data.

In uni-directional mode a detection rate of $\geq 95\%$ with a mutant allele sensitivity of 10% of the normal allele is claimed. Although data solely comprising uni-directional reads has not been assessed in this study (under uni-directional settings/thresholds), a lower sensitivity for uni-directional data is evident from the true uni-directional false negative rate observed for the data in this study (ranging from 0.2-6.9% - table 21). Based on these preliminary findings, we would not recommend analysis reliant solely on automated mutation detection in single orientation reads, but that assessment of a set of high quality uni-directional sequences (tested under uni-directional thresholds) would be a useful follow-up to this study.

One of the quoted strengths of Mutation Surveyor is its ability to de-convolute heterozygote indel mutations into both alleles, as achieved with 89% (155/175) of the indel mutations in our data. This should then permit continuation of the mutational analysis downstream of the frameshift. In our experience Mutation Surveyor was not very successful in automatically detecting mutations after/downstream of an indel deconvolution. Within our data set there were 49 possible opportunities for Mutation Surveyor to detect a substitution mutation downstream of an indel mutation but none of these were detected. Users need to bear in mind that a manufacturer's claims may not always be met.

SoftGenetics® claim that detection sensitivity performs equally well with either terminator or primer chemistries from either gel or capillary systems from all manufacturers of DNA sequencing instrumentation, yet when analysing data generated on a Beckman Coulter there is an increased number of false positive/negative results compared to data produced on other sequencers. In addition read length is not noted by the manufacturer to be a factor in the detection sensitivity, nonetheless we have noted a lower sensitivity in long read sequence data and the optimum fragment size for automated detection appears to be lower than those we tested in our fourth data set.

A major drawback of the software is that, when using the automated mutation analysis, the output table generated has no indication of ROI sequence coverage. Consequently the user is left to check manually the graphical and/or tabulated output to determine if the relevant trace data has been covered and that all the relevant sequence has been covered in both orientations.

Mutation Surveyor is capable of measuring sequence data quality, however the software's definition of 'Low Quality' is set as a default at a very low level. Consequently Mutation Surveyor processes poor quality data indicating a successful analysis for data that would otherwise be rejected by visual inspection. It is clear that the sensitivity of Mutation Surveyor is depressed in data classified as acceptable under the default settings. As it is possible to increase the threshold of quality for acceptance considerable further local validation would be necessary to determine the correct level to match the local sequencing chemistry, read length and electrophoretic separation.

Nevertheless, Mutation Surveyor is a very comprehensive and useful program for detection of mutations in DNA sequence data and can make a very significant contribution in diagnostic laboratories in helping to ease the burden of sequence data analysis. Although we have highlighted weaknesses with the program when it is used in automatic mode with default settings 'out of the box', the user has the facility to alter many parameters which could increase overall sensitivity by tailoring the mutation detection algorithm to local sequencing chemistry, strategy etc. Clearly it is beyond the scope of this Technology Assessment to test all possible configurations. The manufacturer could address these problems by recommending possible configurations of Mutation Surveyor to suit diagnostic laboratories which further minimised the likelihood of false negatives.

Summarised Result	Explanation on Visual Inspection	Numbers of False Results for each Data Set			
		EMQN (1)	CMGS (2)	VariantSEqr (3)	Exon linked (4)
Mutations detected by Mutation Surveyor	-	458	492	249	580
Expected mutations per strand sequenced	-	506	367	41	204
Correctly identified mutations	-	446	335	37	173
False Positives					
	poor quality data	2.0% (10/506)	20.4% (75/367)	517.1% (212/41)	122.1% (249/204)
	Beckman data	0%	22.3% (82/367)	-	-
	true false positives	0.2% (1/506)	0%	0%	77.5% (158/204)
Total		2.2% (11/506)	42.8% (157/367)	517.1%(212/41)	199.5% (407/204)
False Negatives					
Bi-directional false negatives per strand sequenced	poor quality data	0.8% (4/506)	1.1% (4/367)	0%	0%
	low level mosaic	-	-	0%	4.9% (10/204)
	true bi-directional false negative	0%	0.5% (2/367)	4.9% (2/41)	1.0% (2/204)
Uni-directional false negatives	poor quality data	1.2% (6/506)	1.6% (6/367)	0%	0%
	mutation masked by frameshift	9.3% (47/506)	-	2.4% (1/41)	0.5% (1/204)
	loss of resolution	0%	0%	0%	0.5% (1/204)
	mis-aligned data	0%	0.5% (2/367)	0%	1.5% (3/204)
	Beckman data	0%	4.9% (18/367)	-	-
	true uni-directional false negative	0.6% (3/506)	0.2% (1/367)	2.4% (1/41)	6.9% (14/204)
Total		11.9% (60/506)	9.0% (33/367)	9.8% (4/41)	15.2% (31/204)

Table 21. Summarised false positive/negative results from all data sets [% = a false value expressed as a percentage of the total expected number of mutations per strand sequenced for a given data set; true false negatives are classed as those traces which have no immediate visible explanation as to why they were not detected by Mutation Surveyor].

8. ABBREVIATIONS / GLOSSARY

CMGS	Clinical Molecular Genetics Society
cDNA	Complementary DNA
CDS	This designation indicates that the sequence is a coding subsequence derived from a larger sequence.
CFTR	Cystic Fibrosis Transmembrane conductance Regulator
Cx26	Connexin 26, <i>aka:</i> Gap Junction Protein Beta-2; GJB2
Dropping factor	The relative peak intensity drop at the mutation position relative to the neighbouring peaks.
EMQN	European Molecular genetics Quality Network
F/R	Forward/Reverse directions of sequence data
MEN1	Multiple Endocrine Neoplasia, Type I
MS	Mutation Surveyor
Mutation peak height	The maximum intensity of the mutation peak in the mutation electropherogram.
Mutation Score	A figure calculated by Mutation Surveyor indicating the relative likelihood of a mutation being genuine derived from the theoretical calculation of signal to noise ratio with the normal distribution, and the additional two parameters of dropping factor and overlapping factor
NC	Normal Control
NF2	Neurofibromatosis type II
Noise	The median peak height of all smaller mutation peaks (in the mutation electropherogram) within a local section.
Overlapping factor	The indicator of relative shift of the two peaks at the mutation position in the horizontal direction. The overlapping factor calculates the horizontal (time) overlapping percentage of a wild type peak to the mutant peak.
PCR	Polymerase Chain Reaction
ROI	Region of Interest
Signal/Noise ratio	The signal to noise ratio is used to determine the confidence of the peaks, where the confidence is calculated with Gaussian distribution, assuming that the median value (σ) is the noise and the highest value is the signal. The area of the Gaussian curve under 1σ is 68%, 2σ is 95%, and 3σ is 99.7%. The error probability of the mutation peak is $1 - \text{confidence}$.
Trace/Lane Quality Score	A measurement of the average signal to noise ratio, where a lane quality score of 20 signifies that there is 5% noise in that lane. ($n/s = 1/20 = 0.05$). When there is too much noise and a large number of N calls the software is unable to de-convolute and determine the actual sequence trace, Mutation Surveyor defines the trace as being Low Quality.
WT	Wild Type

9. APPENDIX

Minimum System Requirements:

- A 1 GHz Pentium III processor with 128 Mb of RAM
- Microsoft Windows operating systems 2000 / NT / XP

Specification of the PC used to test the program:

- A 2.8 GHz Pentium 4 processor with 1GB of RAM
- Microsoft Windows operating system XP revision 2002

Technology assessments in this series

Automated DNA extraction - Genra Autopure LS	September 2004	http://www.ngri.org.uk/Manchester/Pages/Downloads/GenraHTA/Genra_Autopure_LS_Text_v2.pdf
Applied Biosystems VariantSeqr™ & SeqScape® v2.1 - an assessment using a model system	January 2005	http://www.ngri.org.uk/Manchester/Pages/Downloads/SeqScapeHTA/Vseqr_SeqS_v5.pdf
Gael Quality's Q-Pulse™ v4.2 and Genial Genetic Solutions Ltd Lab Passport TM (Java version) - an assessment of two software systems for quality management	May 2005	http://www.ngri.org.uk/Manchester/Pages/Downloads/Software_HTA_v1.pdf
DNA sequence data analysis - Automated mutation detection using SoftGenetics® Mutation Surveyor™ v2.51	September 2005	http://www.ngri.org.uk/Manchester/Pages/Downloads/MS_HTA_v5.pdf