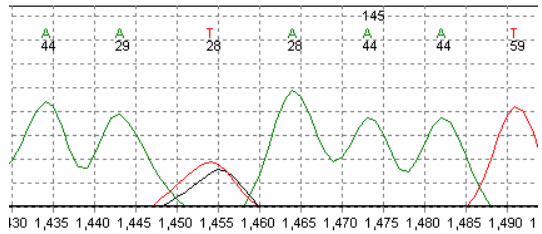
Towards semi-automated analysis of unidirectional sequence data using Mutation Surveyor

Carolyn Tysoe, Beverley Shields, Rebecca Treacy, Shu Yau, Christopher Mattocks, Andrew Wallace and Sian Ellard

On behalf of the CMGS Scientific sub-committee



Outline

Bidirectional *vs* unidirectional sequencing

Semi-automated analysis *vs* visual inspection

CMGS study to assess the sensitivity of Mutation Surveyor

(1) unidirectional data – any missed mutations?

(2) bidirectional data – any missed mutations in one direction?

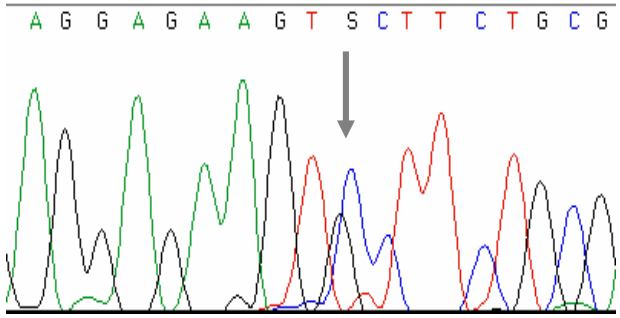
Defining quality parameters for semi-automated analysis

Quality analysis of routine sequencing data (*ABCC8*, n=50)

Conclusions and future work

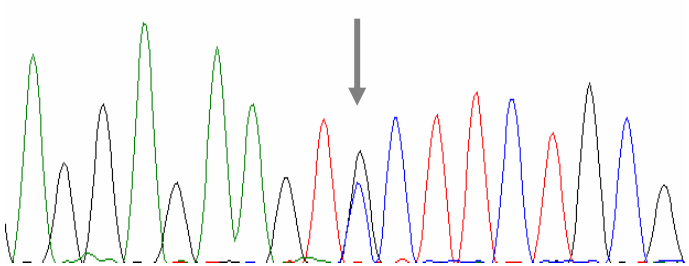
Bidirectional sequencing (2D)
traditionally considered to be the gold standard

1995



Sequencing quality was not perfect > mutations could have been missed

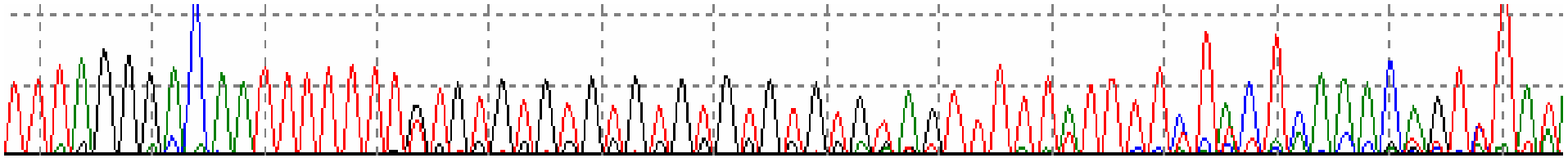
2008



Sequence quality has improved

CMGS Best Practice guidelines recommend 2D
but are out of date (2001) and have been retired

The case for unidirectional (1D) sequencing



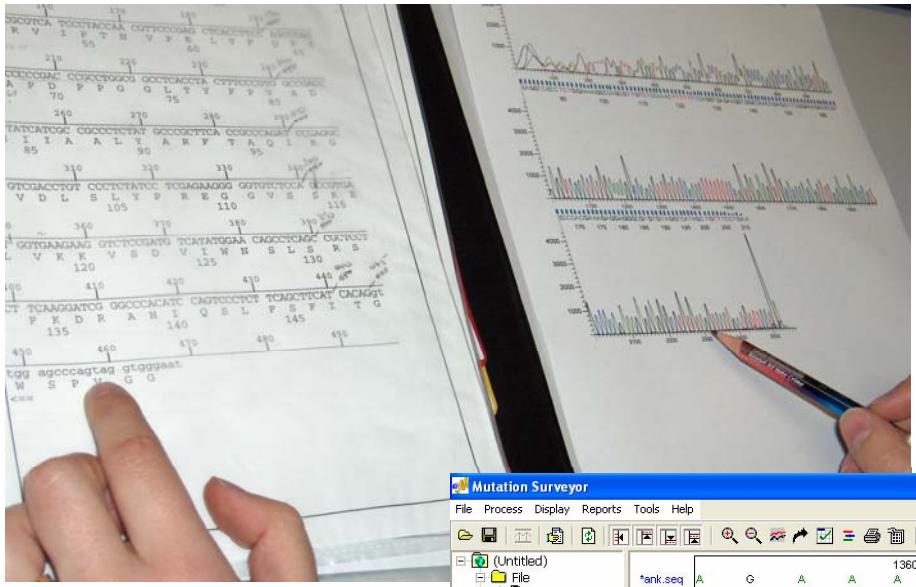
CFTR intron 8

Some regions can only be examined in 1D
eg. intronic insertions/deletions can cause frameshifts

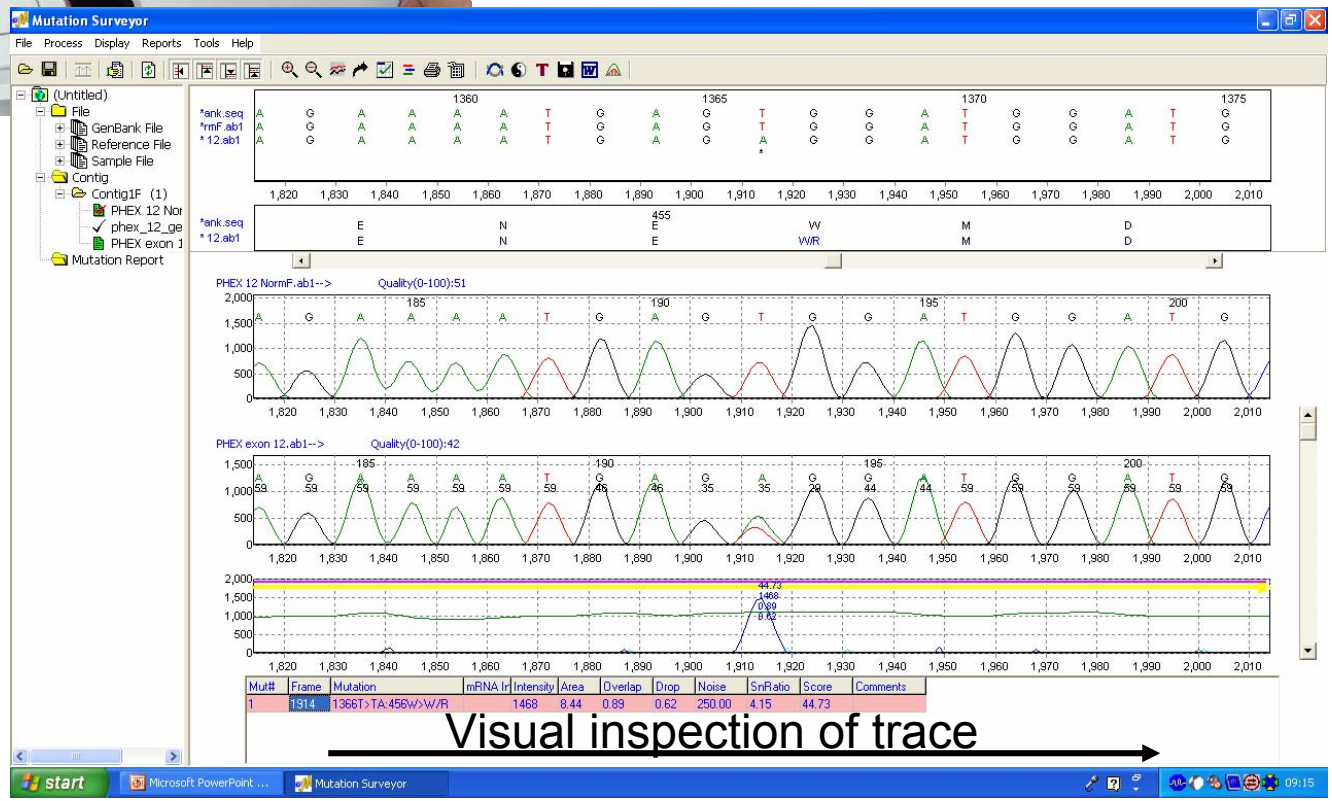
No biological reason for 2D sequencing

SCOBEC has agreed that 1D is acceptable

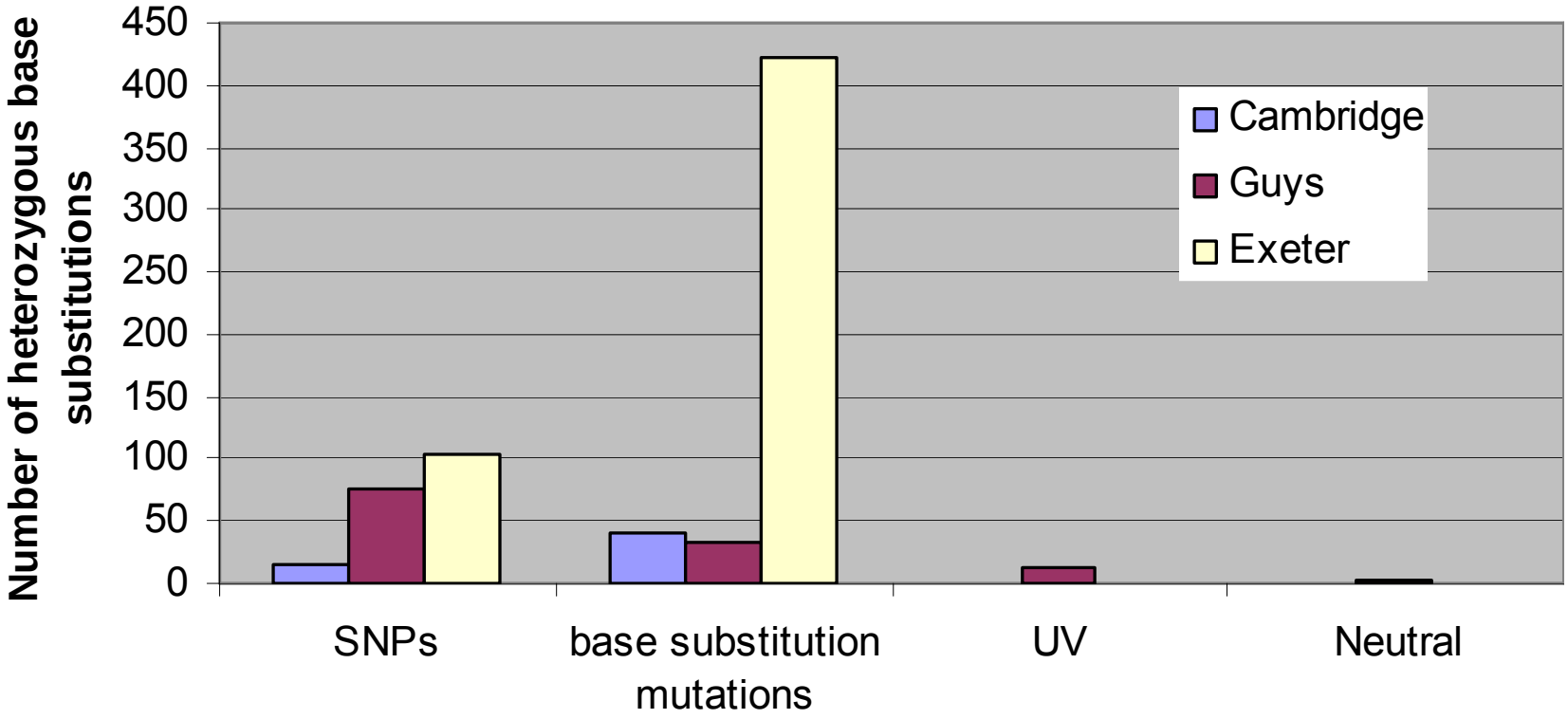
Sequence analysis software has improved



Mutation Surveyor
Annotation and trace comparison



What is the sensitivity of Mutation Surveyor for detecting heterozygous bases in 1D sequencing?



Software detected 701 unique heterozygous base substitutions in 27 genes

No mutations were missed

This data gives us 99.8% confidence that the error rate < 1%

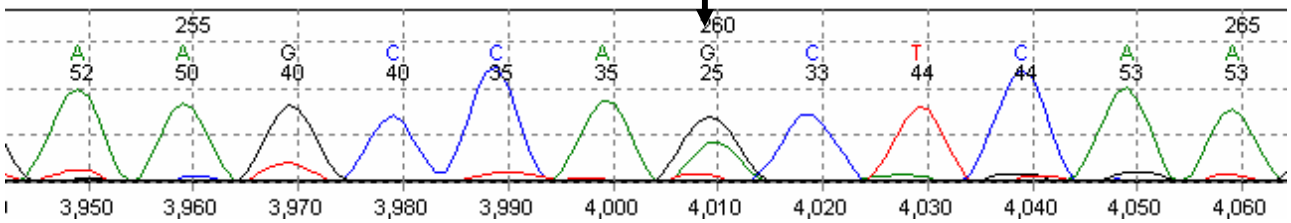
Examples where Mutation Surveyor missed heterozygous bases in one direction during 2D analysis

Detected with 2D settings?

Three examples were provided by CMGS labs

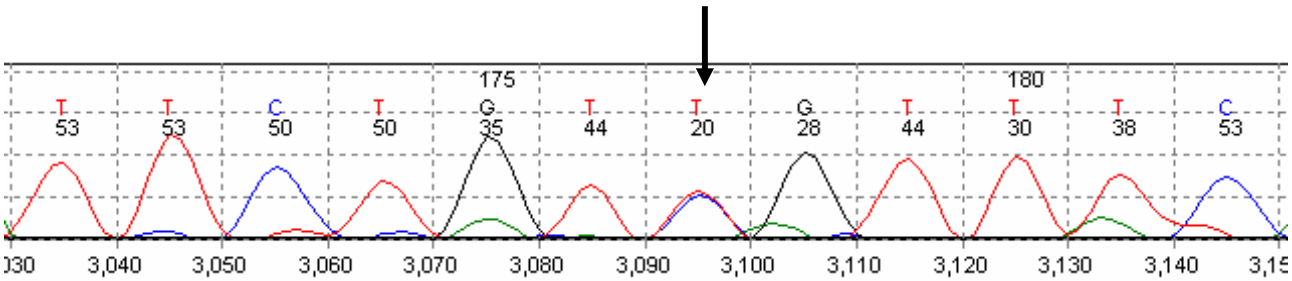
Detected with 1D settings?

No



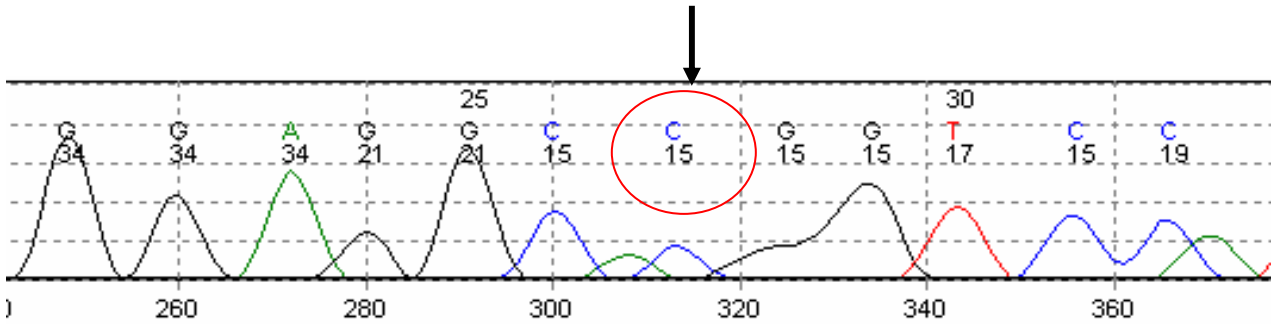
Yes

No



Yes

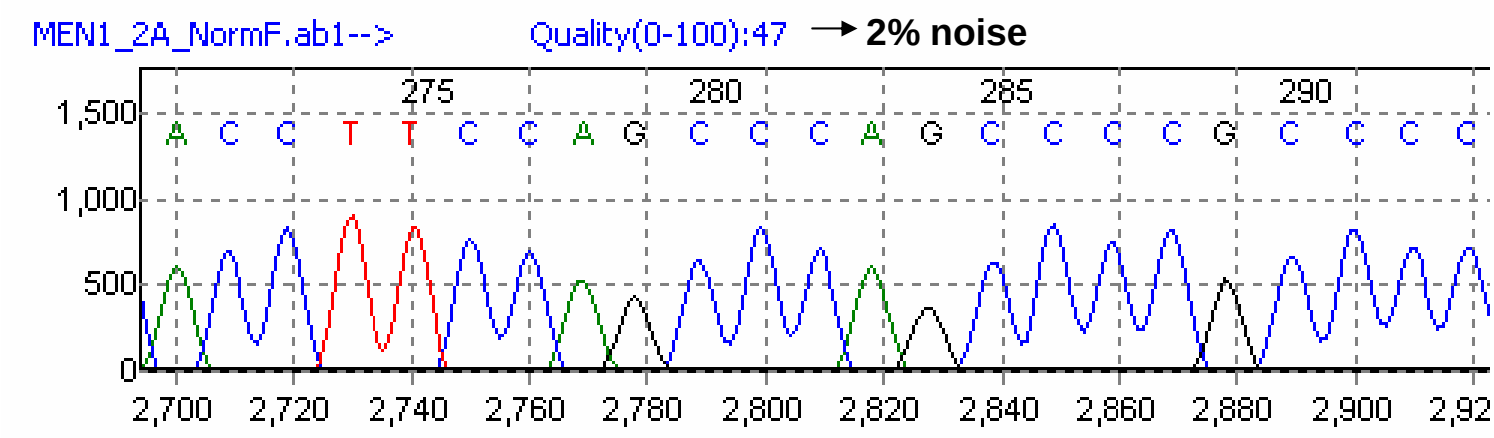
No



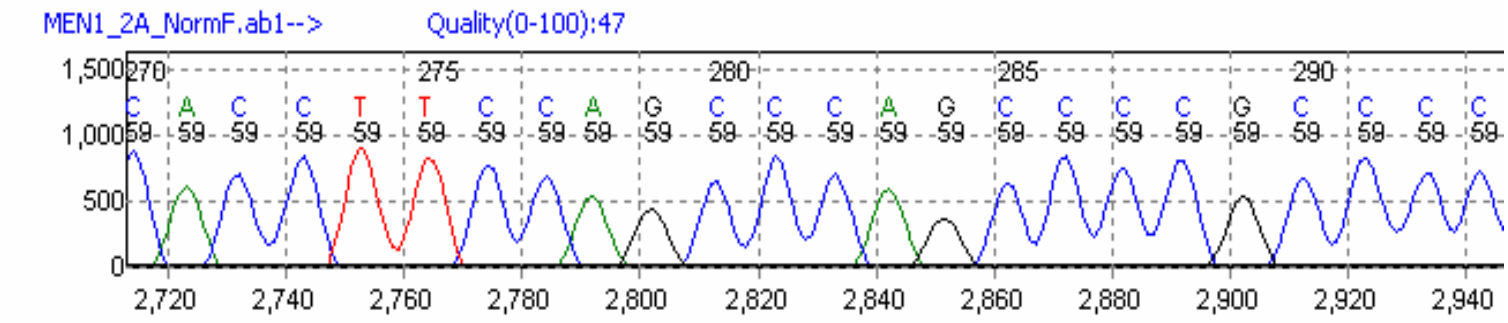
No

How does Mutation Surveyor assess sequence quality?

(1) Quality score - represents signal: noise ratio with an average for the ROI (eg QS 20 = 5% noise)



(2) PHRED scores for each base with an average for the ROI



PHRED scores are numerical values representing the quality of base calling

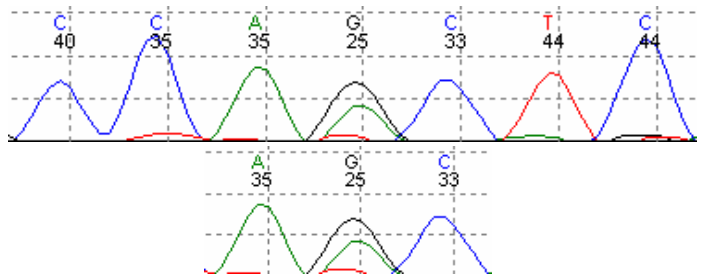
Phred quality score	Probability that the base has been called wrongly	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

MS uses *PHRED-like* scores which are calculated using a modified algorithm

Peak spacing

uncalled/called peak ratio (7 peaks)

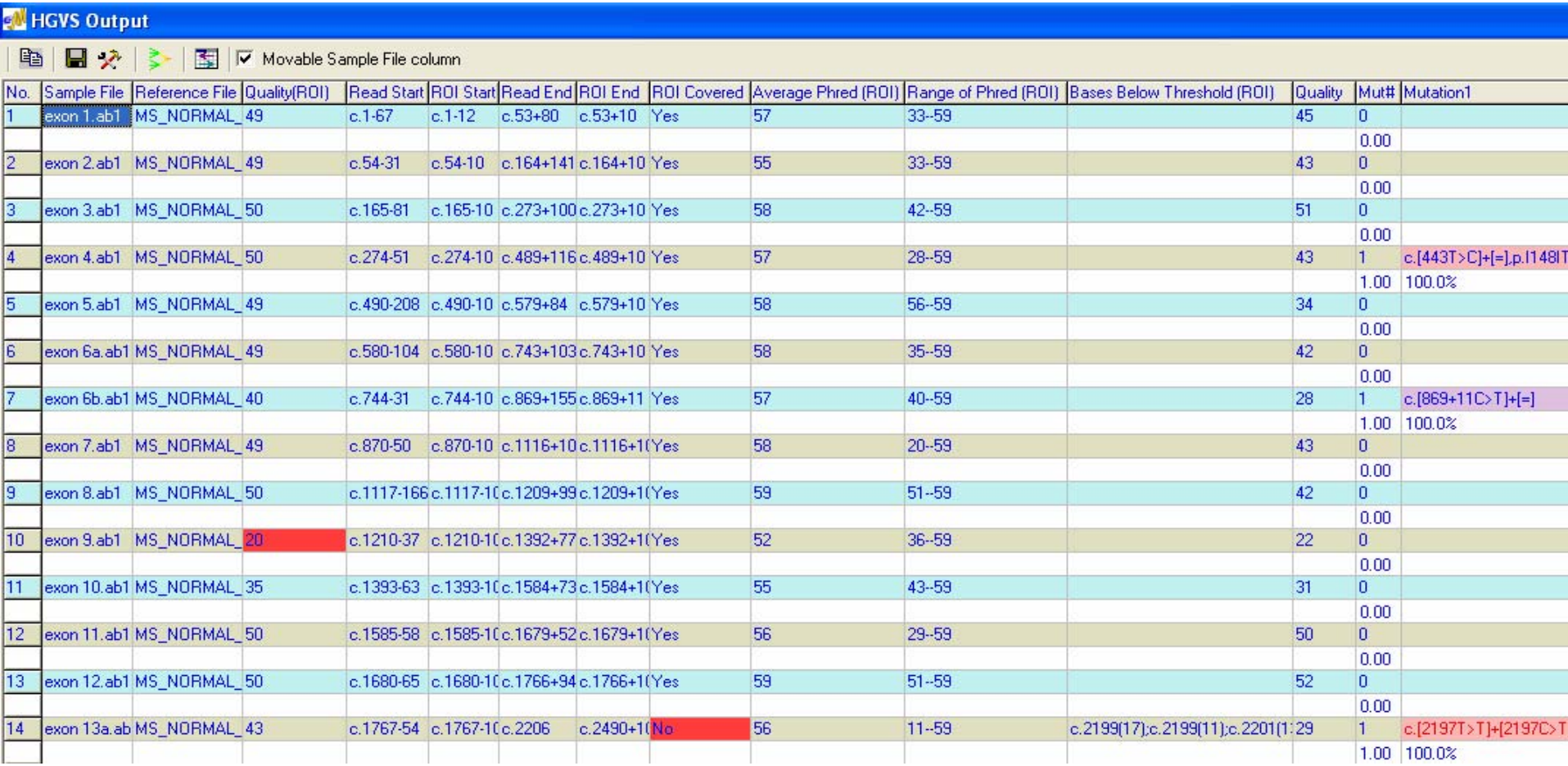
uncalled/called peak ratio (3 peaks)



Peak resolution

Comparable to PHRED except in regions of low quality > may differ by a score of 5

MS quality parameters are displayed in the HGVS table


HGVS Output

Movable Sample File column

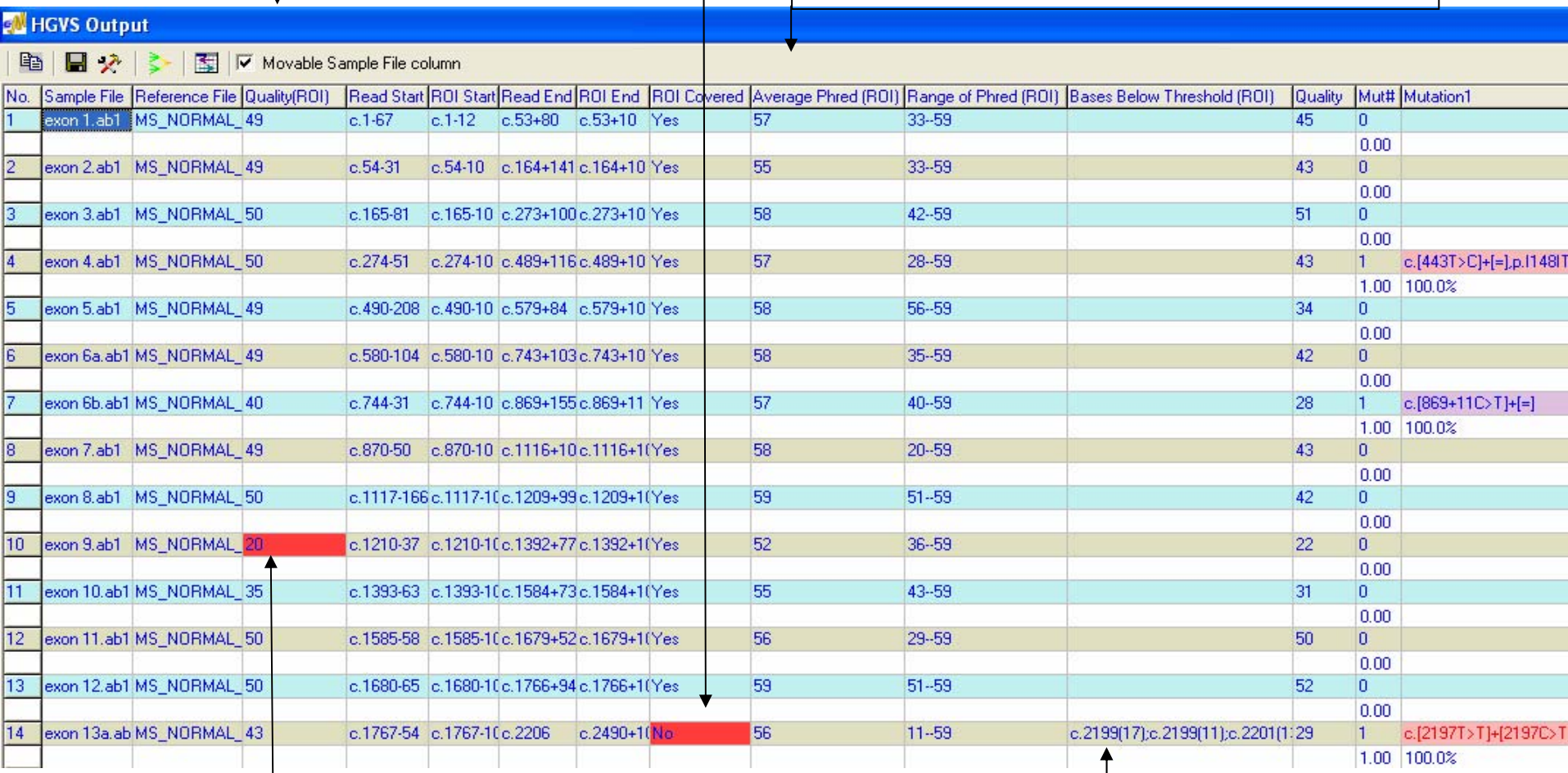
No.	Sample File	Reference File	Quality(ROI)	Read Start	ROI Start	Read End	ROI End	ROI Covered	Average Phred (ROI)	Range of Phred (ROI)	Bases Below Threshold (ROI)	Quality	Mut#	Mutation1
1	exon 1.ab1	MS_NORMAL_	49	c.1-67	c.1-12	c.53+80	c.53+10	Yes	57	33--59		45	0	
													0.00	
2	exon 2.ab1	MS_NORMAL_	49	c.54-31	c.54-10	c.164+141	c.164+10	Yes	55	33--59		43	0	
													0.00	
3	exon 3.ab1	MS_NORMAL_	50	c.165-81	c.165-10	c.273+100	c.273+10	Yes	58	42--59		51	0	
													0.00	
4	exon 4.ab1	MS_NORMAL_	50	c.274-51	c.274-10	c.489+116	c.489+10	Yes	57	28--59		43	1	c.[443T>C]+[=],p.I148[T
													1.00	100.0%
5	exon 5.ab1	MS_NORMAL_	49	c.490-208	c.490-10	c.579+84	c.579+10	Yes	58	56--59		34	0	
													0.00	
6	exon 6a.ab1	MS_NORMAL_	49	c.580-104	c.580-10	c.743+103	c.743+10	Yes	58	35--59		42	0	
													0.00	
7	exon 6b.ab1	MS_NORMAL_	40	c.744-31	c.744-10	c.869+155	c.869+11	Yes	57	40--59		28	1	c.[869+11C>T]+[=]
													1.00	100.0%
8	exon 7.ab1	MS_NORMAL_	49	c.870-50	c.870-10	c.1116+10	c.1116+1(	Yes	58	20--59		43	0	
													0.00	
9	exon 8.ab1	MS_NORMAL_	50	c.1117-166	c.1117-1(	c.1209+99	c.1209+1(	Yes	59	51--59		42	0	
													0.00	
10	exon 9.ab1	MS_NORMAL_	20	c.1210-37	c.1210-1(	c.1392+77	c.1392+1(	Yes	52	36--59		22	0	
													0.00	
11	exon 10.ab1	MS_NORMAL_	35	c.1393-63	c.1393-1(	c.1584+73	c.1584+1(	Yes	55	43--59		31	0	
													0.00	
12	exon 11.ab1	MS_NORMAL_	50	c.1585-58	c.1585-1(	c.1679+52	c.1679+1(	Yes	56	29--59		50	0	
													0.00	
13	exon 12.ab1	MS_NORMAL_	50	c.1680-65	c.1680-1(	c.1766+94	c.1766+1(	Yes	59	51--59		52	0	
													0.00	
14	exon 13a.ab	MS_NORMAL_	43	c.1767-54	c.1767-1(	c.2206	c.2490+1(	No	56	11--59	c.2199(17);c.2199(11);c.2201(1:	29	1	c.[2197T>T]+[2197C>T
													1.00	100.0%

Contains features requested at NHS user meeting

Quality score showing the average signal/noise (S/N) ratio across the region of interest (ROI)

Highlight traces in which the ROI is not covered

Quality score showing the average PHRED-like score across the ROI



No.	Sample File	Reference File	Quality(ROI)	Read Start	ROI Start	Read End	ROI End	ROI Covered	Average Phred (ROI)	Range of Phred (ROI)	Bases Below Threshold (ROI)	Quality	Mut#	Mutation1
1	exon 1.ab1	MS_NORMAL_	49	c.1-67	c.1-12	c.53+80	c.53+10	Yes	57	33--59		45	0	
													0.00	
2	exon 2.ab1	MS_NORMAL_	49	c.54-31	c.54-10	c.164+141	c.164+10	Yes	55	33--59		43	0	
													0.00	
3	exon 3.ab1	MS_NORMAL_	50	c.165-81	c.165-10	c.273+100	c.273+10	Yes	58	42--59		51	0	
													0.00	
4	exon 4.ab1	MS_NORMAL_	50	c.274-51	c.274-10	c.489+116	c.489+10	Yes	57	28--59		43	1	c.[443T>C]+[=],p.I148T
													1.00	100.0%
5	exon 5.ab1	MS_NORMAL_	49	c.490-208	c.490-10	c.579+84	c.579+10	Yes	58	56--59		34	0	
													0.00	
6	exon 6a.ab1	MS_NORMAL_	49	c.580-104	c.580-10	c.743+103	c.743+10	Yes	58	35--59		42	0	
													0.00	
7	exon 6b.ab1	MS_NORMAL_	40	c.744-31	c.744-10	c.869+155	c.869+11	Yes	57	40--59		28	1	c.[869+11C>T]+[=]
													1.00	100.0%
8	exon 7.ab1	MS_NORMAL_	49	c.870-50	c.870-10	c.1116+10	c.1116+1(	Yes	58	20--59		43	0	
													0.00	
9	exon 8.ab1	MS_NORMAL_	50	c.1117-166	c.1117-1(	c.1209+99	c.1209+1(	Yes	59	51--59		42	0	
													0.00	
10	exon 9.ab1	MS_NORMAL_	20	c.1210-37	c.1210-1(	c.1392+77	c.1392+1(	Yes	52	36--59		22	0	
													0.00	
11	exon 10.ab1	MS_NORMAL_	35	c.1393-63	c.1393-1(	c.1584+73	c.1584+1(	Yes	55	43--59		31	0	
													0.00	
12	exon 11.ab1	MS_NORMAL_	50	c.1585-58	c.1585-1(	c.1679+52	c.1679+1(	Yes	56	29--59		50	0	
													0.00	
13	exon 12.ab1	MS_NORMAL_	50	c.1680-65	c.1680-1(	c.1766+94	c.1766+1(	Yes	59	51--59		52	0	
													0.00	
14	exon 13a.ab	MS_NORMAL_	43	c.1767-54	c.1767-1(	c.2206	c.2490+1(	No	56	11--59	c.2199(17);c.2199(11);c.2201(1	29	1	c.[2197T>T]+[2197C>T
													1.00	100.0%

Highlight traces which fall below a user defined quality score

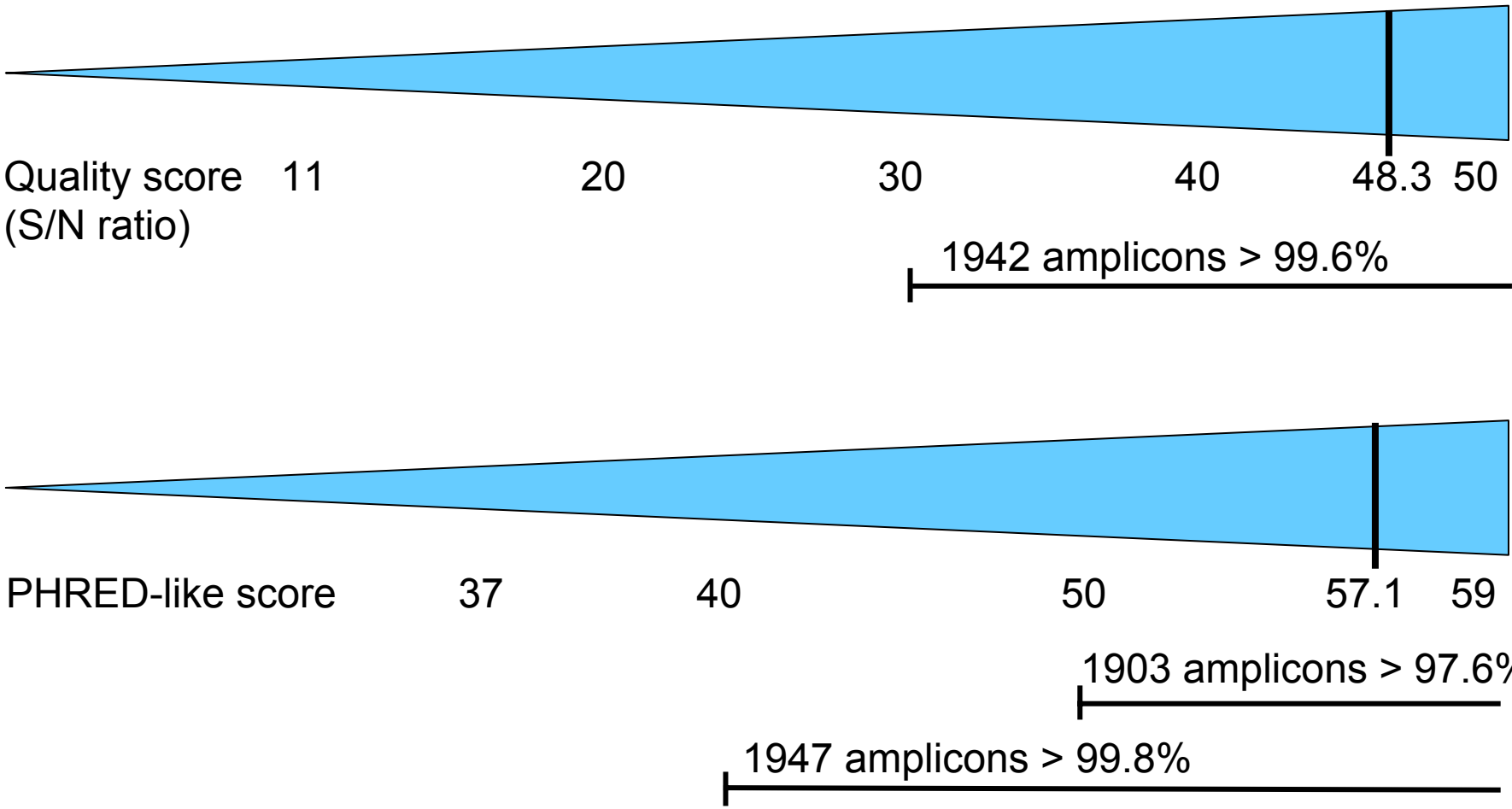
Highlight bases within the ROI that do not meet a user-defined quality score

Sequencing data for *ABCC8* gene in a consecutive series of 50 unselected patients

ABCC8 gene has 39 exons

50 patients = 1950 amplicons

Distribution of mean quality scores (ROI) within 1950 amplicons



97.6% amplicons have a QS ≥ 30 or PHRED-like score ≥ 50

Distribution of quality scores for individual bases within the region of interest (ROI)

1950 amplicons

322,100 bases within ROI had a visual inspection

How many bases within the ROI are low quality? (have a PHRED-like score ≤ 20)

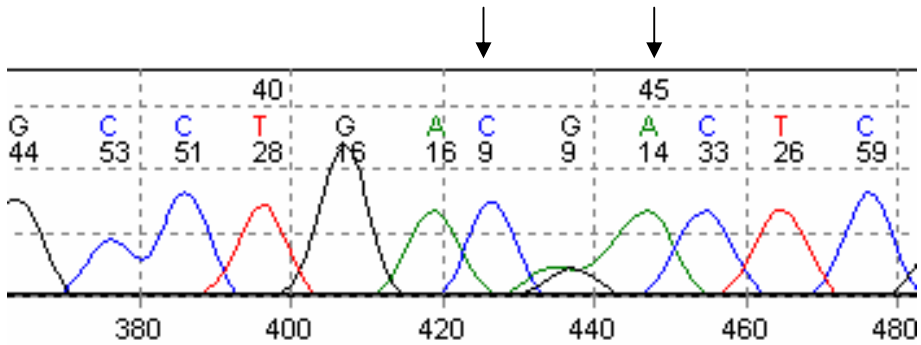
419 / 322,100 = 0.13% bases had PHRED-like score ≤ 20 and would need a visual inspection

2/3 of bases with PHRED <20 are within the context of a heterozygous base

143 poor quality bases > need visual check (0.05%)

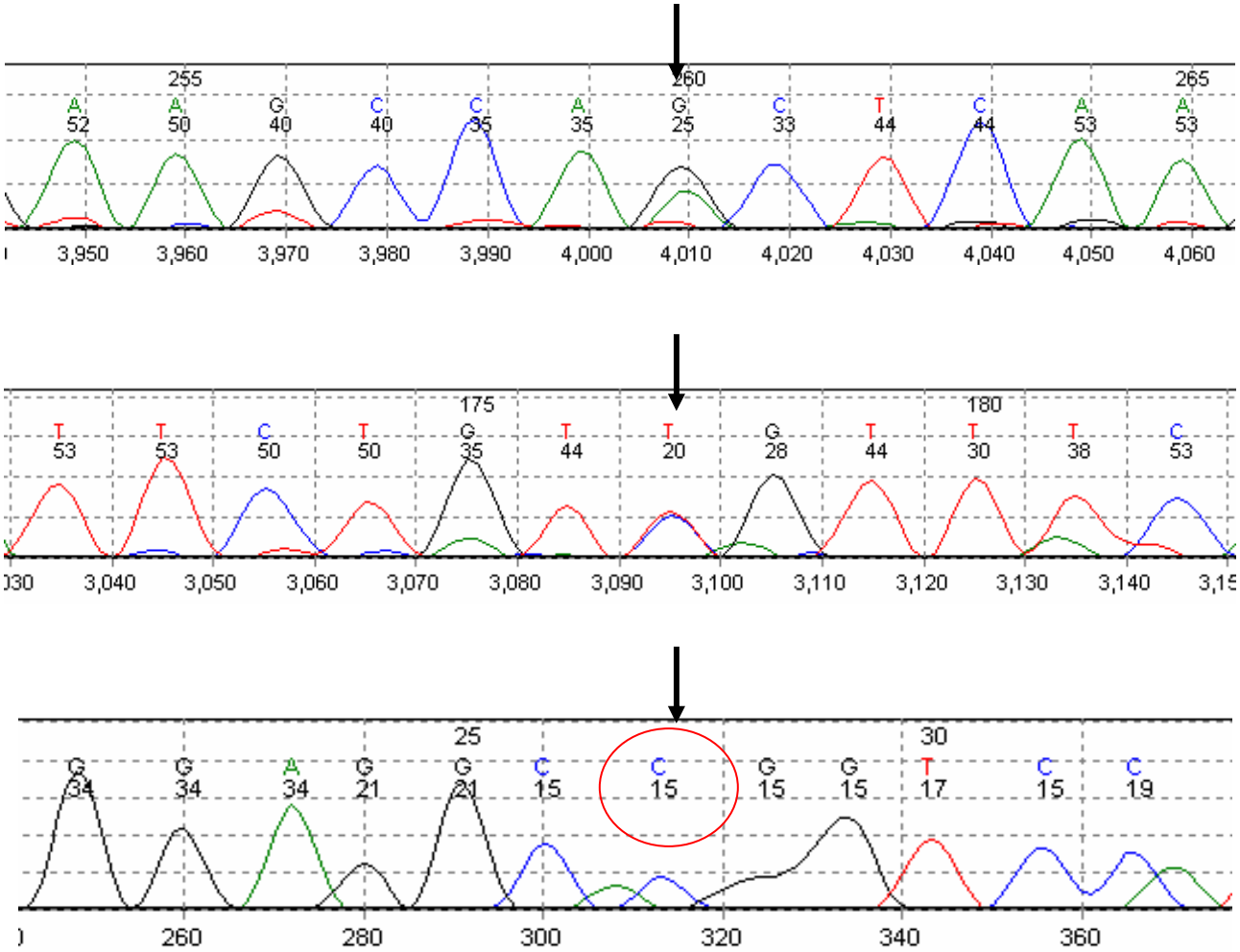
419 / 322,100 = 0.13% bases had PHRED-like score ≤ 20
and would need a visual inspection

266 are in the context of a heterozygous base



Bases flanking SNPs may have low PHRED scores

Missed heterozygous base has a PHRED-like score < 20



Conclusions

(1) Unidirectional sequence analysis using Mutation Surveyor has a high sensitivity (99.8% confidence that error rate <1%)

(2) All 3 heterozygous base substitutions missed in one direction by 2D analysis could be either detected using 1D settings or highlighted as low quality bases

(3) Analysis of a consecutive series of 1950 *ABCC8* amplicons showed that 97.6% of sequences had a mean PHRED-like score ≥50 or a quality score ≥30 for the ROI

Towards semi-automated sequence analysis

A visual inspection would be required for:

- Sequences containing mutations or polymorphisms (from Mutation report)
- Low quality bases (ideally by link from HGVS table)
- Sequences with ROI quality scores or average PHRED-like scores that fall below a threshold

Further work is required across multiple laboratories in order to establish appropriate thresholds for visual inspection

↓

Development of new CMGS Best Practice Guidelines to include semi-automated analysis of sequence data

CMGS CLINICAL MOLECULAR GENETICS SOCIETY

Acknowledgements

The CMGS labs for providing data

All the Exeter team - Dr. Ann-Marie Patch, Michael Day and Piers Fulton (ABCC8 data)

Dr. Michael Weedon

SoftGenetics