Detection and Quantification of Sequence Variants from Sanger Sequencing Traces

Determination of minor alleles by analyzing peak height data

The introduction of semi-automated fluorescent dye-terminator DNA Sequencing using capillary electrophoresis (aka CE or Sanger sequencing) has revolutionized life and medical sciences by unraveling complete genomes and the elucidation of genetic structures of many organisms. The primary information and value of the DNA sequencing process is the identification of the nucleotides and of possible sequence variants. A largely unknown and unexplored feature of fluorescent Sanger sequencing traces is the quantitative information embedded therein. With the growing need for quantifying somatic mutations in tumor tissue, emerging mutations in viral genomes conferring drug resistance, or the amount of methylation in a particular CpG locus, it is desirable to exploit the potential of the quantitative information obtained from sequencing traces.

In this application note we review two freely available software applications that help to extract and present the peak height data of Sanger sequencing traces for quantitative data analysis.

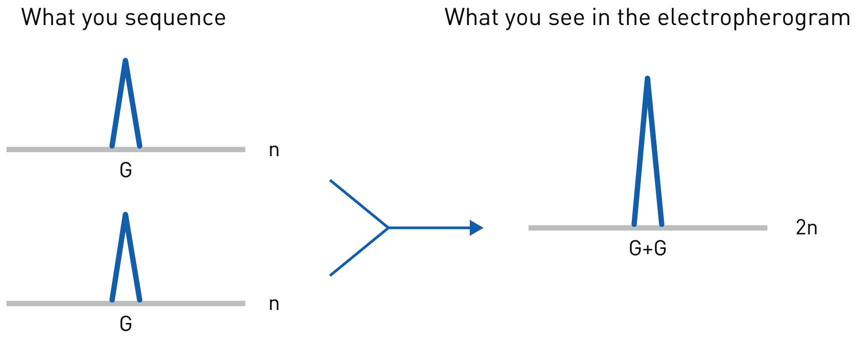


Figure 1: Homozygosity. The peak signal is the sum of the peak signals from the two haploid input DNAs.

The composite electropherogram and the challenge of mixed basecalling

DNA basecalling software programs analyze fluorescent Sanger sequencing traces and reveal the base identities of a DNA sample along with quality values (phred scores) which indicate the reliability of the basecall. In a typical PCR-based sequencing project that uses DNA from a diploid organism both copies of an allele are sequenced simultaneously. Compared to the hypothetical signal n resulting from a single allele, the observed signal is actually the result of both alleles combined, or 2n.

An individual peak in a sequencing trace representing a homozygous base is a composite mixture of two identical bases each contributing approximately half of the fluorescent signal in relative fluorescent units (RFU) to a given peak height (see Figure 1). Hence, the loss (e.g., by amplification drop out) of one allele will typically lead to a drop of signal by half (illustrated in Figure 2).

In the case of heterozygosity at an allele the resulting peak pair migrates at the same or

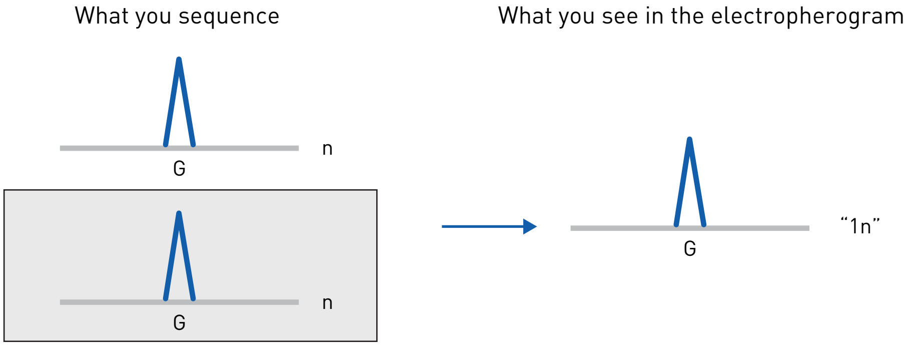


Figure 2: Allele Drop-Out. The peak signal is only approximately half of the signal expected in the case of homozygosity.

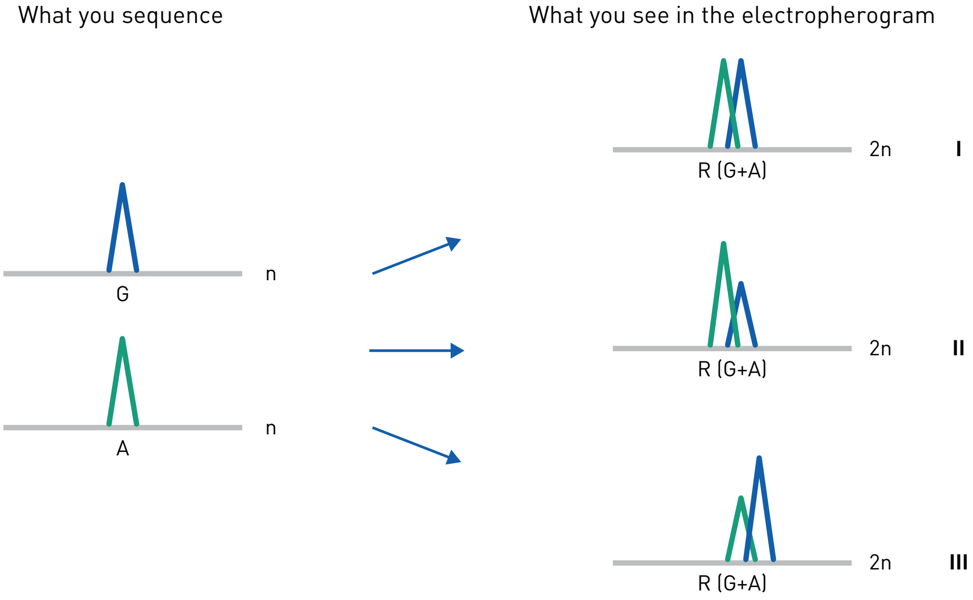


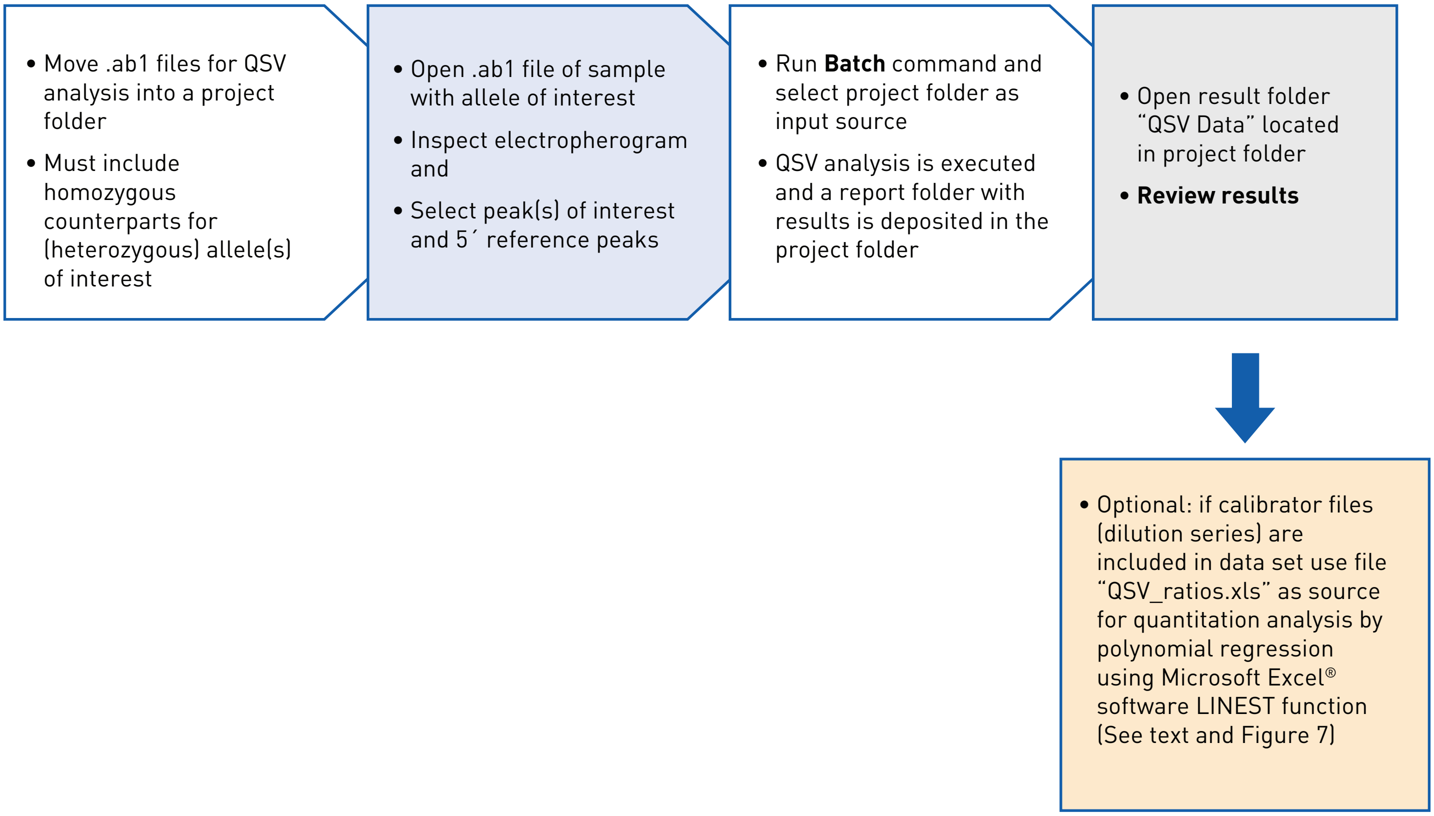
Figure 3: Heterozygosity: sequencing a heterozygous allele may ideally present in an electropherogram as a balanced peak pair (Outcome I) or may appear somewhat imbalanced (Outcome II or III). The specific outcome for a given peak pair is typically highly reproducible and depends on the local sequence context.

similar position as a mixed base. The signal strength of each component is approximately half of the homozygous counterpart. Ideally, the two heterozygous peaks appear to be of equal height (see outcome I) but in reality they may occur somewhat unbalanced (outcome II or III) depending on the DNA strand sequenced and sequence-dependent context. This complicates the determination of peak height ratios. However, this imbalance phenomenon is typically highly reproducible for a given allele from sample to sample and can be accounted for using homozygous control samples (see text). (Figure 3)

The simple principle that the proportion of each of two sequence variants in a mixture determine the relative heights of the peaks that represent each variant in a sequence electropherogram has inspired Ian Carr and colleagues from the University of Leeds Institute of Molecular Medicine to develop a software application that exploits the quantitative information embedded in a sequencing trace.

Homozygous and heterozygous sequence variants are readily detected by commercial and public domain sequence analysis software packages. However, minor sequence variants such as they are found in somatic mutations in tumor tissue or in emerging mutations in subpopulations of microbial or viral organisms often elude detection because the abundance of the minor allele is too low for triggering a (mixed) basecall.

The heights of the primary and secondary peaks in a mixed-base situation are the most important attributes for basecalling. If the peak height ratio of a secondary to a



- Move .ab1 files for QSV analysis into a project folder
- Must include homozygous counterparts for (heterozygous) allele(s) of interest

- Open .ab1 file of sample with allele of interest
- Inspect electropherogram and
- Select peak(s) of interest and 5´ reference peaks

- Run **Batch** command and select project folder as input source
- QSV analysis is executed and a report folder with results is deposited in the project folder

- Open result folder “QSV Data” located in project folder
- **Review results**

- Optional: if calibrator files (dilution series) are included in data set use file “QSV_ratios.xls” as source for quantitation analysis by polynomial regression using Microsoft Excel® software LINEST function (See text and Figure 7)

Figure 4: Overview of the QSVanalyzer workflow.

primary peak drops below 30% (or other user-set threshold) it is usually not considered and therefore not called out as a mixed base.

In this application note we will review the paper and the QSVanalyzer software published by Ian Carr et al. from the University of Leeds, UK and recommend its utility for the detection and quantification of sequence variants.

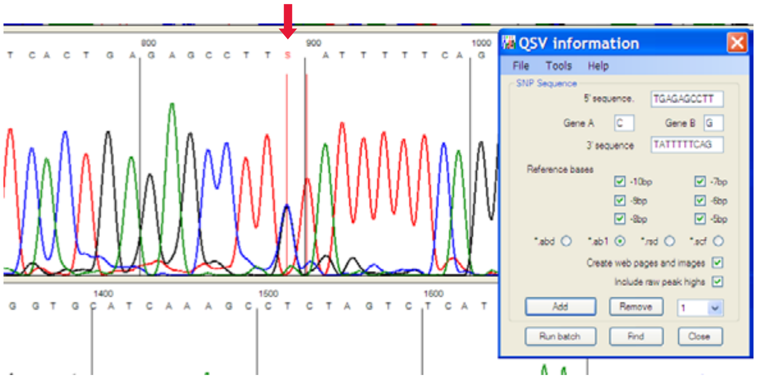
We also describe a new bioinformatics utility, ab1PeakReporter, which is available on the Life Technologies web site. The utility provides numerical peak height data of Sanger sequencing traces allowing the quantitative analysis of peak height data. To that end, we show how minor alleles can be quantified by polynomial regression analysis using Microsoft Excel® software.

Inferring Allelic Variant Ratios using QSVAnalyzer

In 2009, Carr et al. published a paper describing the QSVanalyzer desktop application in the journal Bioinformatics. QSVanalyzer enables the high-throughput quantification of the proportions of DNA sequences containing single-nucleotide sequence variants (SNVs) from fluorescent Sanger sequencing traces. The paper is open access and can be downloaded with supplementary data from [1]. The QSVanalyzer application including original sequencing trace files used in the study can be downloaded from http://dna.leeds.ac.uk/qsv/ .

In the paper, Carr et al. demonstrated the utility of the method for estimation of copy number proportions (CNPs) for various quantitative sequence variant (QSV) types such as common regular SNPs, paralogous sequence variants (PSV) and SNPs in the background of copy number variation (CNV).

An important concept presented in the paper is the normalization of electropherograms: Fluorescent dideoxynucleotide terminators are incorporated dependent on their sequence context and may appear imbalanced in heterozygous mixed bases (see Figure 3). Further, the amount of template DNA and other factors affect the absolute peak height. Therefore, relative (rather than absolute) peak heights are determined by comparing the variant nucleotide’s peak height to that of an invariant nucleotide located 5’ (upstream) where one can assume a neutral sequence background, i.e., no variant–introduced effects. The software also corrects for the background baseline signal in each



800
900
1000
T C A C T G A G A G C C T T S A T T T T T C A G
1400
1500
1600
G G T G C A T C A A A G C C T C T A G T C T C A T
QSV information
File
Tools
Help
SNP Sequence
Add
Remove
Run batch
Find
Close

5' sequence. TGAGAGCCTT

Gene A C Gene B G

3' sequence TATTTTTCAG

Reference bases

☑ -10bp ☑ -7bp
☑ -9bp ☑ -6bp
☑ -8bp ☑ -5bp

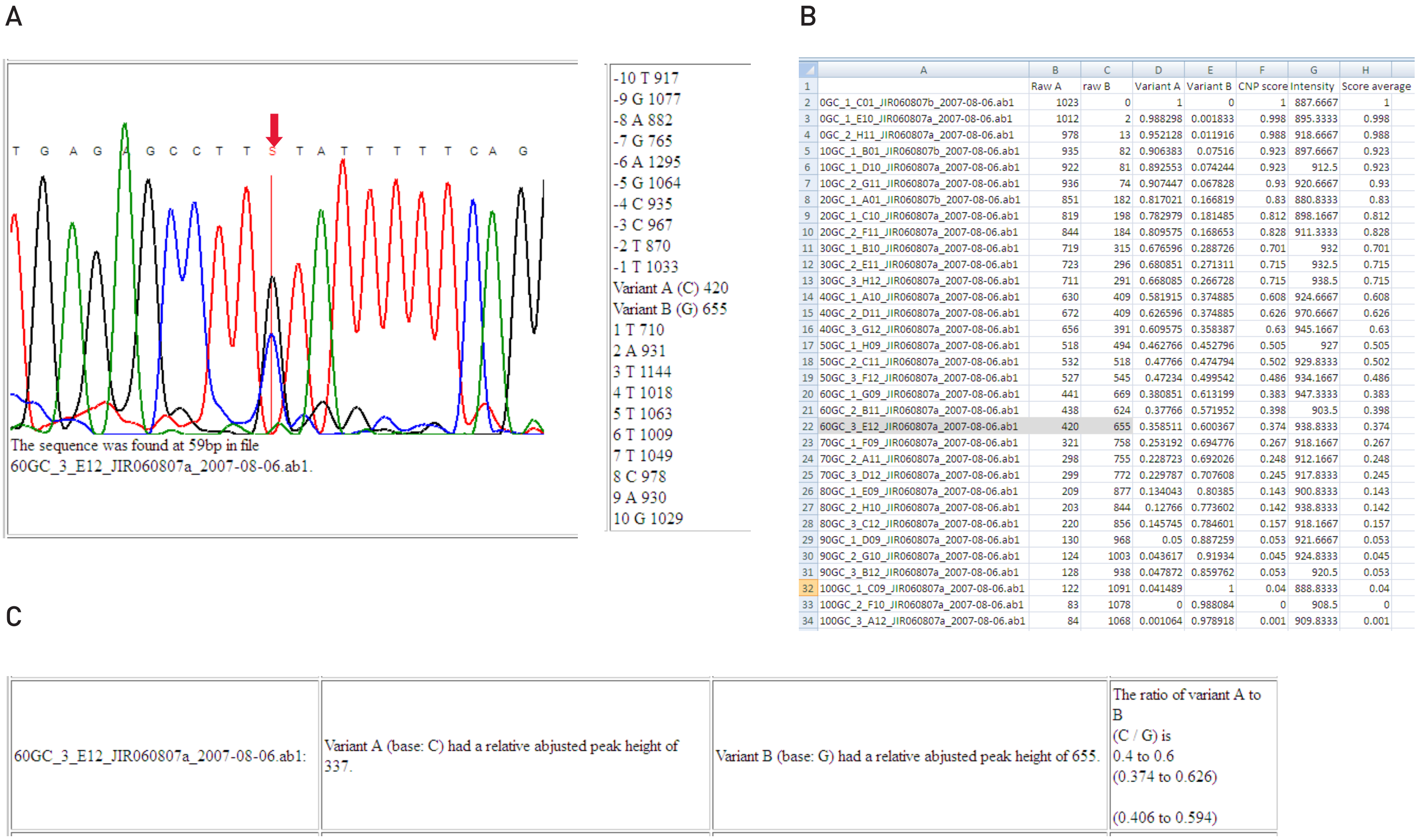
*.abd ○ *.ab1 ● *.rsd ○ *.scf ○

Create web pages and images ☑

Include raw peak highs ☑

1

Figure 5: Electropherogram viewer with a heterozygous allele "S (for C and G)" near scan # 900.



The sequence was found at 59bp in file 60GC_3_E12_JIR060807a_2007-08-06.ab1.

-10 T 917
-9 G 1077
-8 A 882
-7 G 765
-6 A 1295
-5 G 1064
-4 C 935
-3 C 967
-2 T 870
-1 T 1033
Variant A (C) 420
Variant B (G) 655
1 T 710
2 A 931
3 T 1144
4 T 1018
5 T 1063
6 T 1009
7 T 1049
8 C 978
9 A 930
10 G 1029

	A	B	C	D	E	F	G	H
1		Raw A	raw B	Variant A	Variant B	CNP score	Intensity	Score average
2	0GC_1_C01_JIR060807b_2007-08-06.ab1	1023	0	1	0	1	887.6667	1
3	0GC_1_E10_JIR060807a_2007-08-06.ab1	1012	2	0.988298	0.001833	0.998	895.3333	0.998
4	0GC_2_H11_JIR060807a_2007-08-06.ab1	978	13	0.952128	0.011916	0.988	918.6667	0.988
5	10GC_1_B01_JIR060807b_2007-08-06.ab1	935	82	0.906383	0.07516	0.923	897.6667	0.923
6	10GC_1_D10_JIR060807a_2007-08-06.ab1	922	81	0.892553	0.074244	0.923	912.5	0.923
7	10GC_2_G11_JIR060807a_2007-08-06.ab1	936	74	0.907447	0.067828	0.93	920.6667	0.93
8	20GC_1_A01_JIR060807b_2007-08-06.ab1	851	182	0.817021	0.166819	0.83	880.8333	0.83
9	20GC_1_C10_JIR060807a_2007-08-06.ab1	819	198	0.782979	0.181485	0.812	898.1667	0.812
10	20GC_2_F11_JIR060807a_2007-08-06.ab1	844	184	0.809575	0.168653	0.828	911.3333	0.828
11	30GC_1_B10_JIR060807a_2007-08-06.ab1	719	315	0.676596	0.288726	0.701	932	0.701
12	30GC_2_E11_JIR060807a_2007-08-06.ab1	723	296	0.680851	0.271311	0.715	932.5	0.715
13	30GC_3_H12_JIR060807a_2007-08-06.ab1	711	291	0.668085	0.266728	0.715	938.5	0.715
14	40GC_1_A10_JIR060807a_2007-08-06.ab1	630	409	0.581915	0.374885	0.608	924.6667	0.608
15	40GC_2_D11_JIR060807a_2007-08-06.ab1	672	409	0.626596	0.374885	0.626	970.6667	0.626
16	40GC_3_G12_JIR060807a_2007-08-06.ab1	656	391	0.609575	0.358387	0.63	945.1667	0.63
17	50GC_1_H09_JIR060807a_2007-08-06.ab1	518	494	0.462766	0.452796	0.505	927	0.505
18	50GC_2_C11_JIR060807a_2007-08-06.ab1	532	518	0.47766	0.474794	0.502	929.8333	0.502
19	50GC_3_F12_JIR060807a_2007-08-06.ab1	527	545	0.47234	0.499542	0.486	934.1667	0.486
20	60GC_1_G09_JIR060807a_2007-08-06.ab1	441	669	0.380851	0.613199	0.383	947.3333	0.383
21	60GC_2_B11_JIR060807a_2007-08-06.ab1	438	624	0.37766	0.571952	0.398	903.5	0.398
22	60GC_3_E12_JIR060807a_2007-08-06.ab1	420	655	0.358511	0.600367	0.374	938.8333	0.374
23	70GC_1_F09_JIR060807a_2007-08-06.ab1	321	758	0.253192	0.694776	0.267	918.1667	0.267
24	70GC_2_A11_JIR060807a_2007-08-06.ab1	298	755	0.228723	0.692026	0.248	912.1667	0.248
25	70GC_3_D12_JIR060807a_2007-08-06.ab1	299	772	0.229787	0.707608	0.245	917.8333	0.245
26	80GC_1_E09_JIR060807a_2007-08-06.ab1	209	877	0.134043	0.80385	0.143	900.8333	0.143
27	80GC_2_H10_JIR060807a_2007-08-06.ab1	203	844	0.12766	0.773602	0.142	938.8333	0.142
28	80GC_3_C12_JIR060807a_2007-08-06.ab1	220	856	0.145745	0.784601	0.157	918.1667	0.157
29	90GC_1_D09_JIR060807a_2007-08-06.ab1	130	968	0.05	0.887259	0.053	921.6667	0.053
30	90GC_2_G10_JIR060807a_2007-08-06.ab1	124	1003	0.043617	0.91934	0.045	924.8333	0.045
31	90GC_3_B12_JIR060807a_2007-08-06.ab1	128	938	0.047872	0.859762	0.053	920.5	0.053
32	100GC_1_C09_JIR060807a_2007-08-06.ab1	122	1091	0.041489	1	0.04	888.8333	0.04
33	100GC_2_F10_JIR060807a_2007-08-06.ab1	83	1078	0	0.988084	0	908.5	0
34	100GC_3_A12_JIR060807a_2007-08-06.ab1	84	1068	0.001064	0.978918	0.001	909.8333	0.001

60GC_3_E12_JIR060807a_2007-08-06.ab1:	Variant A (base: C) had a relative abjusted peak height of 337.	Variant B (base: G) had a relative abjusted peak height of 655.	The ratio of variant A to B (C / G) is 0.4 to 0.6 (0.374 to 0.626) (0.406 to 0.594)

Figure 6: Output reports of the QSVanalyzer application. (**A**) Widget of the electropherogram accompanied by peak heights of the area. (**B**) Comprehensive Excel-readable table with raw and reference-adjusted data. (**C**) Final Quantitative Sequence Variant (QSV) report with adjusted peak heights (see Carr et al. for details).

trace and subtracts the allele-specific “background noise” from the relative peak height for a final normalized peak height (NPH). To calculate the QSV ratio, the program needs two reference sequences, each containing the homozygous allele of the two variants.

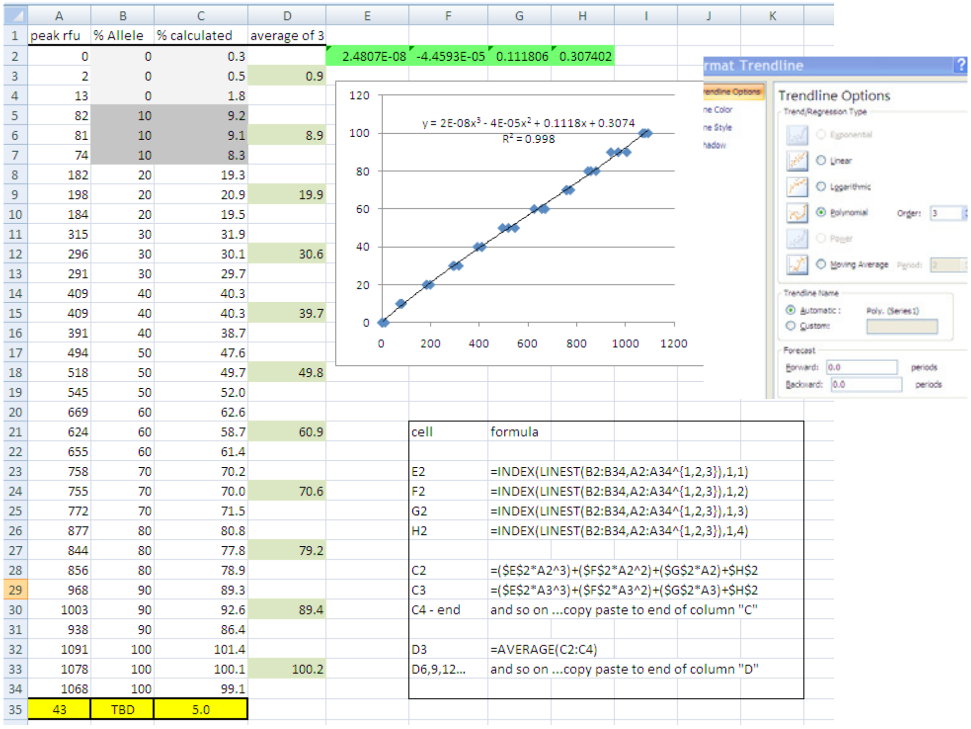
A detailed, illustrated guide for use of the application can be found on http://dna.leeds.ac.uk/qsv/guide/ and a set of original .ab1 sample files with a differential dilution series is available on http://dna.leeds.ac.uk/qsv/download.

Figure 5 shows the electropherogram viewer with a heterozygous allele “S (for C and G)” near scan # 900 along with the QSV information setup window where this and other alleles of interest are selected for subsequent batch analysis of QSV ratios (step 2 of workflow).

Figure 6 shows QSVanalyzer results for a mixed DNA sample that contained a pre-mixed ratio of 40% variant A (C nucleotide blue trace) and 60% variant B (G nucleotide black trace). QSVanalyzer reported a ratio of 0.4 to 0.6 for A:B (see report 6C).

What is the limit of detection for minor alleles?

The Carr paper provides a web link to sets of original sequencing files from three dilution series each consisting of 11 samples in triplicates with nominal sequence variant proportions 10:0, 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9, and 0:10. We have used these sequencing traces to ask whether it would be possible to detect a minority allele at 10% and distinguish it significantly from background noise. The QSVanalyzer application was used to process the data set “G in CG” provided by the authors and the output Table shown in Figure 6B was opened with Microsoft Excel® software.



	A	B	C	D	E	F	G	H
1	peak rfu	% Allele	% calculated	average of 3				
2	0	0	0.3		2.4807E-08	-4.4593E-05	0.111806	0.307402
3	2	0	0.5	0.9				
4	13	0	1.8					
5	82	10	9.2					
6	81	10	9.1	8.9				
7	74	10	8.3					
8	182	20	19.3					
9	198	20	20.9	19.9				
10	184	20	19.5					
11	315	30	31.9					
12	296	30	30.1	30.6				
13	291	30	29.7					
14	409	40	40.3					
15	409	40	40.3	39.7				
16	391	40	38.7					
17	494	50	47.6					
18	518	50	49.7	49.8				
19	545	50	52.0					
20	669	60	62.6					
21	624	60	58.7	60.9				
22	655	60	61.4					
23	758	70	70.2					
24	755	70	70.0	70.6				
25	772	70	71.5					
26	877	80	80.8					
27	844	80	77.8	79.2				
28	856	80	78.9					
29	968	90	89.3					
30	1003	90	92.6	89.4				
31	938	90	86.4					
32	1091	100	101.4					
33	1078	100	100.1	100.2				
34	1068	100	99.1					
35	43	TBD	5.0					

cell	formula
E2	=INDEX(LINEST(B2:B34,A2:A34^{1,2,3}),1,1)
F2	=INDEX(LINEST(B2:B34,A2:A34^{1,2,3}),1,2)
G2	=INDEX(LINEST(B2:B34,A2:A34^{1,2,3}),1,3)
H2	=INDEX(LINEST(B2:B34,A2:A34^{1,2,3}),1,4)
C2	=($E$2*A2^3)+($F$2*A2^2)+($G$2*A2)+$H$2
C3	=($E$2*A3^3)+($F$2*A3^2)+($G$2*A3)+$H$2
C4 - end	and so on ...copy paste to end of column "C"
D3	=AVERAGE(C2:C4)
D6,9,12...	and so on ...copy paste to end of column "D"

rmat Trendline

Trendline Options

Trend/Regression Type: ● Polynomial, Order: 3

Trendline Name: ● Automatic : Poly. (Series1)

Forecast: Forward: 0.0 periods; Backward: 0.0 periods

Figure 7: A simple polynomial equation calculator using the LINEST function in Microsoft Excel® allows the estimation of allele proportion in % (column C) in relation to peak height in RFU (column A).

To that end we entered the “raw” peak height data for variant B (see Figure 6B column C) into column A of a new spreadsheet and entered the admixed values of this particular allele (0, 10, 20%, etc.) into column B. Next, we applied a scatter plot of the data and used the trend line function in Excel with a polynomial of order 3 for curve fitting. This operation typically yielded a good correlation coefficient (>0.98). We also checked the “Display equation on chart” box which shows the components of the polynomial in the graph. Next we applied the LINEST function in Excel to solve the polynomial equation so that we can calculate for a given peak height (measured as RFU) value the corresponding amount of allele proportion (column C). The required formulas and steps for this are shown in the box in Figure 7. By entering an RFU value into cell A35 the estimated allele proportion is obtained: in the case of 43 RFU it is 5%. Since in this particular experiment and the given allele the background noise is around 5 RFU (average of 0, 2, 13 RFU), it is conceivable that the peak for a minority allele of 5% proportion is potentially detectable.

Note the excellent correlation of averaged measured values (column D) with the theoretical proportions of allele amount in this dilution series (column B). The peak height of an (hypothetical or real) allele of interest at 43 RFU entered into cell A35 was calculated to correspond to an allele of 5% which may be distinguishable from background (approximately 5 RFU; see cells A2–A4).

Taken together, quantification of minor alleles in the 5–10% range may be feasible for at least one allele of an allele pair provided that the experimental system is sufficiently supported with replicates and controlled with calibrator samples. Sequencing and data analysis of the opposite DNA strand may provide further information and resolution.

The ab1PeakReporter utility provides quantitative information of fluorescent Sanger sequencing peak traces
Numerical data describing the raw and processed sequencing traces are embedded in the .ab1 file but are not readily visible using common sequence analysis software. The architecture of the .ab1 file is described in detail in a white paper [2].

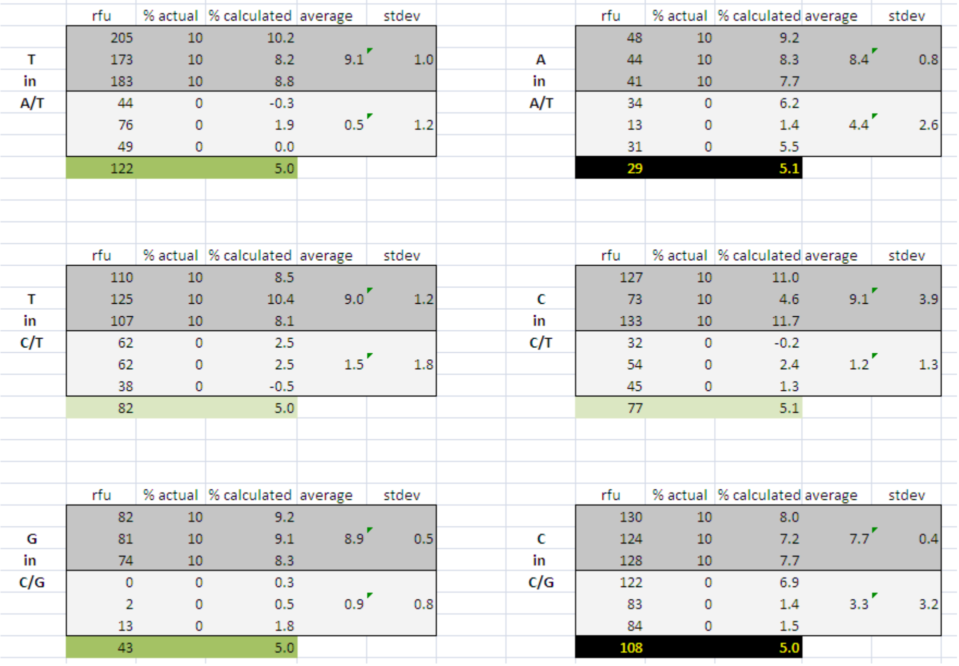
To meet the need for quantitative information from Sanger sequencing traces we have developed a basic utility that reads an .ab1 sequencing file and exports the trace data in various numerical formats.

The ab1PeakReporter utility can be accessed via https://apps.lifetechnologies.com/ab1peakreporter

(Logging in to your Life Technologies customer account is required to use the tool.)

The ab1PeakReporter tool extracts and presents the numerical information from Sanger sequencing traces into an Excel-readable file so that base peak characteristics (reflecting, e.g., allele proportions) can be studied quantitatively using downstream software such as spreadsheet processors.

A batch of up to 96 .ab1 sequencing files can be uploaded into the ab1PeakReporter tool and processed, then exported as a zip file back to a local drive. The zip



	rfu	% actual	% calculated	average	stdev
	205	10	10.2		
T	173	10	8.2	9.1	1.0
in	183	10	8.8		
A/T	44	0	-0.3		
	76	0	1.9	0.5	1.2
	49	0	0.0		
	122		5.0		

	rfu	% actual	% calculated	average	stdev
	48	10	9.2		
A	44	10	8.3	8.4	0.8
in	41	10	7.7		
A/T	34	0	6.2		
	13	0	1.4	4.4	2.6
	31	0	5.5		
	29		5.1		

	rfu	% actual	% calculated	average	stdev
	110	10	8.5		
T	125	10	10.4	9.0	1.2
in	107	10	8.1		
C/T	62	0	2.5		
	62	0	2.5	1.5	1.8
	38	0	-0.5		
	82		5.0		

	rfu	% actual	% calculated	average	stdev
	127	10	11.0		
C	73	10	4.6	9.1	3.9
in	133	10	11.7		
C/T	32	0	-0.2		
	54	0	2.4	1.2	1.3
	45	0	1.3		
	77		5.1		

	rfu	% actual	% calculated	average	stdev
	82	10	9.2		
G	81	10	9.1	8.9	0.5
in	74	10	8.3		
C/G	0	0	0.3		
	2	0	0.5	0.9	0.8
	13	0	1.8		
	43		5.0		

	rfu	% actual	% calculated	average	stdev
	130	10	8.0		
C	124	10	7.2	7.7	0.4
in	128	10	7.7		
C/G	122	0	6.9		
	83	0	1.4	3.3	3.2
	84	0	1.5		
	108		5.0		

Figure 8: Data from polynomial regression analysis of peak height data of a particular allele containing defined proportions; only values for 0% and 10% are listed (dilution series data provided by Carr et al. 2009). RFU = relative fluorescent units = peak height.

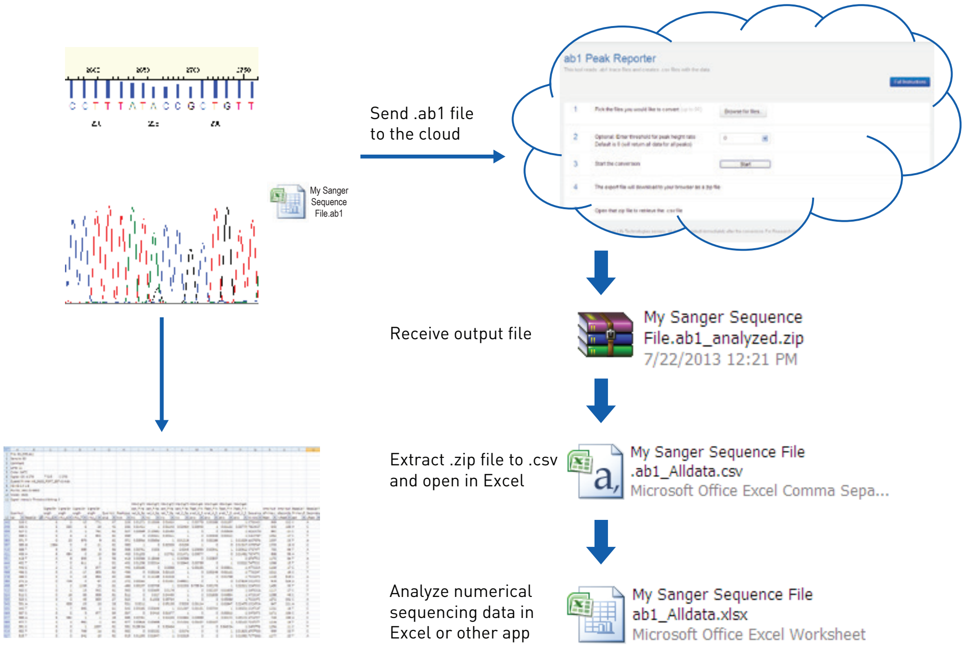
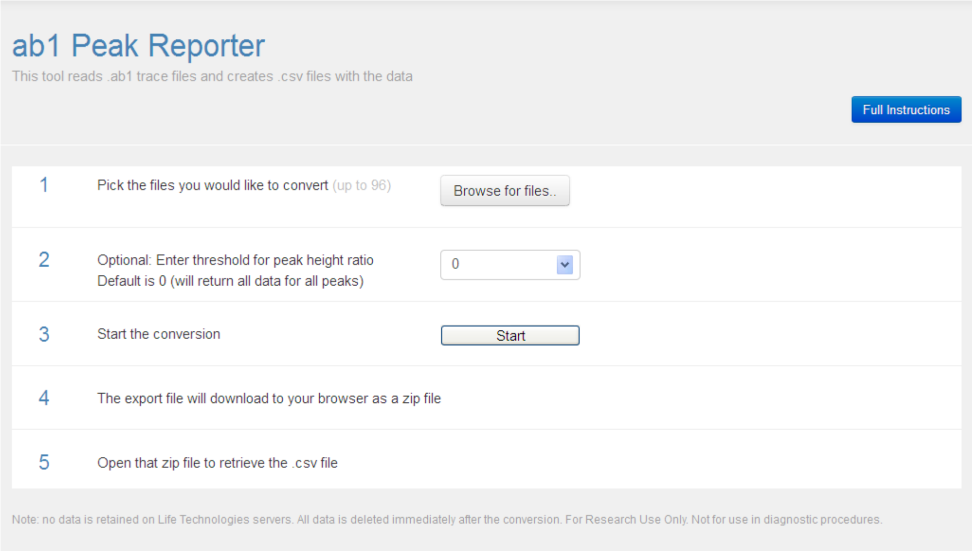

My Sanger Sequence File.ab1
Send .ab1 file to the cloud
ab1 Peak Reporter
Start the conversion
Start
Receive output file
My Sanger Sequence File.ab1_analyzed.zip
7/22/2013 12:21 PM
Extract .zip file to .csv and open in Excel
My Sanger Sequence File .ab1_Alldata.csv
Microsoft Office Excel Comma Sepa...
Analyze numerical sequencing data in Excel or other app
My Sanger Sequence File ab1_Alldata.xlsx
Microsoft Office Excel Worksheet

Figure 9: The ab1PeakReporter workflow.

file is extracted and opened as an Excel-readable .csv file.

The tool is very simple to use:

1. Browse for .ab1 sample files (up to 96) to upload.
2. (Optional) Enter a value between 0–100 to set a detection threshold for a secondary peak; a default value of "0" will detect all peaks including background; a value of "5" will detect and list all secondary peaks over 5% in addition to the primary peak.
3. Select the sample files to analyze. Up to 96 files can be processed at a time.
4. Click the Start button.
5. You will be prompted to open or save a zip folder with the analyzed results to a location on a local computer drive.
6. Extracting the zip folder will yield .csv files that can be opened using Microsoft Excel® or other spreadsheet processing software.
7. Use the data for customized downstream analysis such as the determination of allele ratios (e.g., methylated vs. un-methylated CpG residues in bisulfite-converted DNA) or quantification of minor alleles.



ab1 Peak Reporter

This tool reads .ab1 trace files and creates .csv files with the data

1 Pick the files you would like to convert (up to 96)

2 Optional: Enter threshold for peak height ratio
Default is 0 (will return all data for all peaks)

0

3 Start the conversion

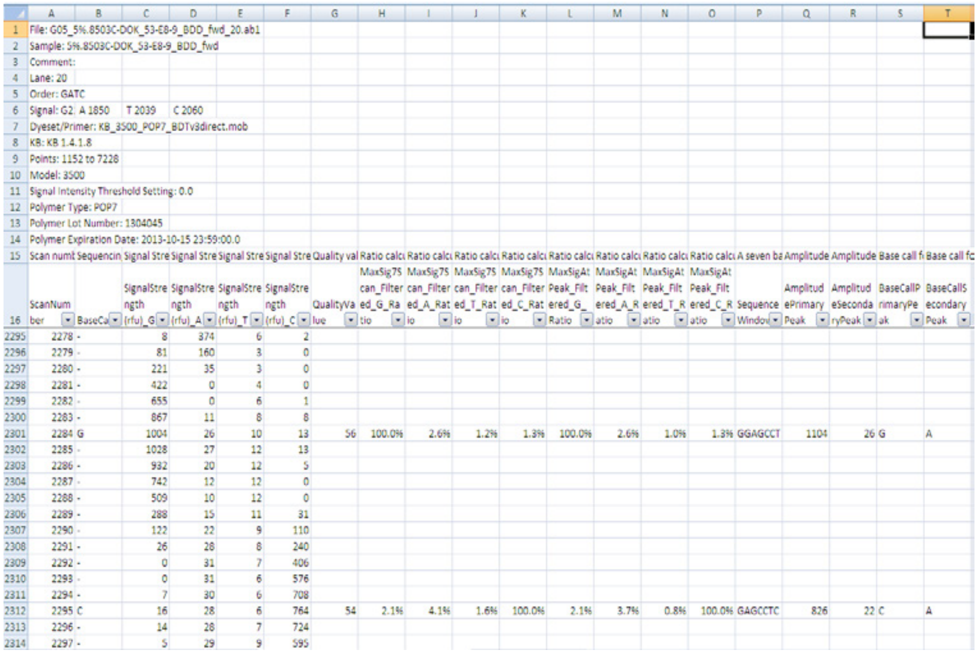
4 The export file will download to your browser as a zip file

5 Open that zip file to retrieve the .csv file

Note: no data is retained on Life Technologies servers. All data is deleted immediately after the conversion. For Research Use Only. Not for use in diagnostic procedures.

Figure 10: User interface of the .ab1Tracer application.

The (your sample name here) _Alldata.csv file (Figure 11) lists all peak height values of all 4 nucleotide traces at all scans along with primary base identification at the location of the amplitude. Rows 1–16 contain a header with basic sample file and run information.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	File: G05_5%.8503C-DOK_53-E8-9_BDD_fwd_20.ab1																			
2	Sample: 5%.8503C-DOK_53-E8-9_BDD_fwd																			
3	Comment:																			
4	Lane: 20																			
5	Order: GATC																			
6	Signal: G2	A 1850	T 2039	C 2060																
7	Dyeset/Primer: KB_3500_POP7_BDTv3direct.mob																			
8	KB: KB 1.4.1.8																			
9	Points: 1152 to 7228																			
10	Model: 3500																			
11	Signal Intensity Threshold Setting: 0.0																			
12	Polymer Type: POP7																			
13	Polymer Lot Number: 1304045																			
14	Polymer Expiration Date: 2013-10-15 23:59:00.0																			
15	Scan numt	Sequencin	Signal Stre	Signal Stre	Signal Stre	Signal Stre	Quality val	Ratio calc	Ratio calc	Ratio calc	Ratio calc	Ratio calc	Ratio calc	Ratio calc	Ratio calc	A seven ba	Amplitude	Amplitude	Base call f	Base call fo
16	ScanNumber	BaseCa	SignalStrength (rfu)_G	SignalStrength (rfu)_A	SignalStrength (rfu)_T	SignalStrength (rfu)_C	QualityValue	MaxSig7Scan_Filtered_G_Ratio	MaxSig7Scan_Filtered_A_Ratio	MaxSig7Scan_Filtered_T_Ratio	MaxSig7Scan_Filtered_C_Ratio	MaxSigAtPeak_Filtered_G_Ratio	MaxSigAtPeak_Filtered_A_Ratio	MaxSigAtPeak_Filtered_T_Ratio	MaxSigAtPeak_Filtered_C_Ratio	SequenceWindov	AmplitudePrimaryPeak	AmplitudeSecondaryPeak	BaseCallPrimaryPeak	BaseCallSecondaryPeak
2295	2278	-	8	374	6	2														
2296	2279	-	81	160	3	0														
2297	2280	-	221	35	3	0														
2298	2281	-	422	0	4	0														
2299	2282	-	655	0	6	1														
2300	2283	-	867	11	8	8														
2301	2284	G	1004	26	10	13	56	100.0%	2.6%	1.2%	1.3%	100.0%	2.6%	1.0%	1.3%	GGAGCCT	1104	26	G	A
2302	2285	-	1028	27	12	13														
2303	2286	-	932	20	12	5														
2304	2287	-	742	12	12	0														
2305	2288	-	509	10	12	0														
2306	2289	-	288	15	11	31														
2307	2290	-	122	22	9	110														
2308	2291	-	26	28	8	240														
2309	2292	-	0	31	7	406														
2310	2293	-	0	31	6	576														
2311	2294	-	7	30	6	708														
2312	2295	C	16	28	6	764	54	2.1%	4.1%	1.6%	100.0%	2.1%	3.7%	0.8%	100.0%	GAGCCTC	826	22	C	A
2313	2296	-	14	28	7	724														
2314	2297	-	5	29	9	595														

Figure 11: The comprehensive ab1PeakReporter AllData Table (_Alldata.csv file).

Below row 16 the following components are listed:

Column A: the scan number

Column B: primary peak as identified by KB™ Basecaller

Columns C–F: continuous peak heights for nucleotide traces G, A, T, C, respectively

Column G: the phred Quality value is shown

Re-create the electropherogram plot in Excel

The electropherogram can be generated using the line graph plot function in Excel by selecting cells A–F or B–F, then go to tab "Insert" and select "Chart > Line" (Figure 12). The electropherogram aids in visual interpretation of ambiguous loci, assisting in distinguishing genuine peaks from background noise or artifacts.

Applying filters aids in data exploration

The next step is to filter out the uncalled scans; this will enable customized display of data and is a great aid in exploring the data. To set filters click on row 16, go to tab "Data" and select "Filter" (Figure 13).

The Filter tool allows selective display of data by (un-)checking individual data points, sorting, and various ranges and rank formats in its "Number Filters" section.

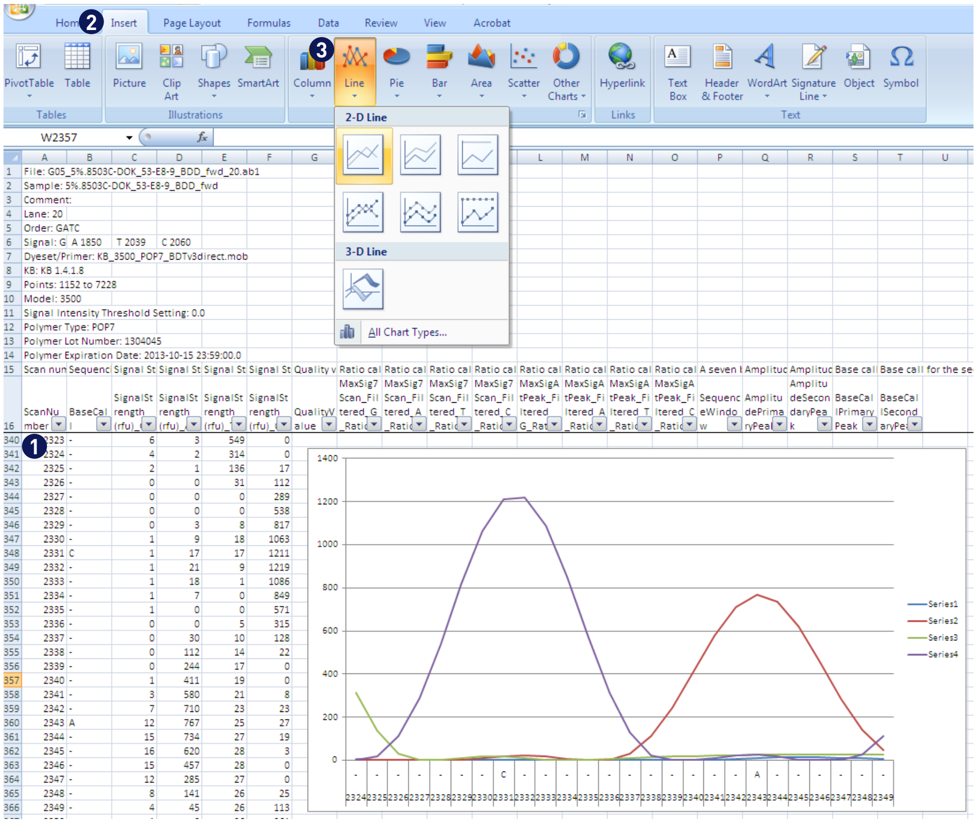
Condensing the table to basecalled data only

Using the Filter tool, the table can be condensed to display only the basecalled data points. This feature is useful for transferring data into a database for subsequent archiving, and further exploratory or statistical analysis.

To condense the data table to basecalled data points only, 1) Click the Filter icon in column B, "BaseCall", 2) Uncheck the "–" box, 3) click the "OK" button (Figure 14).

Find loci of interest with "Sequence Window"

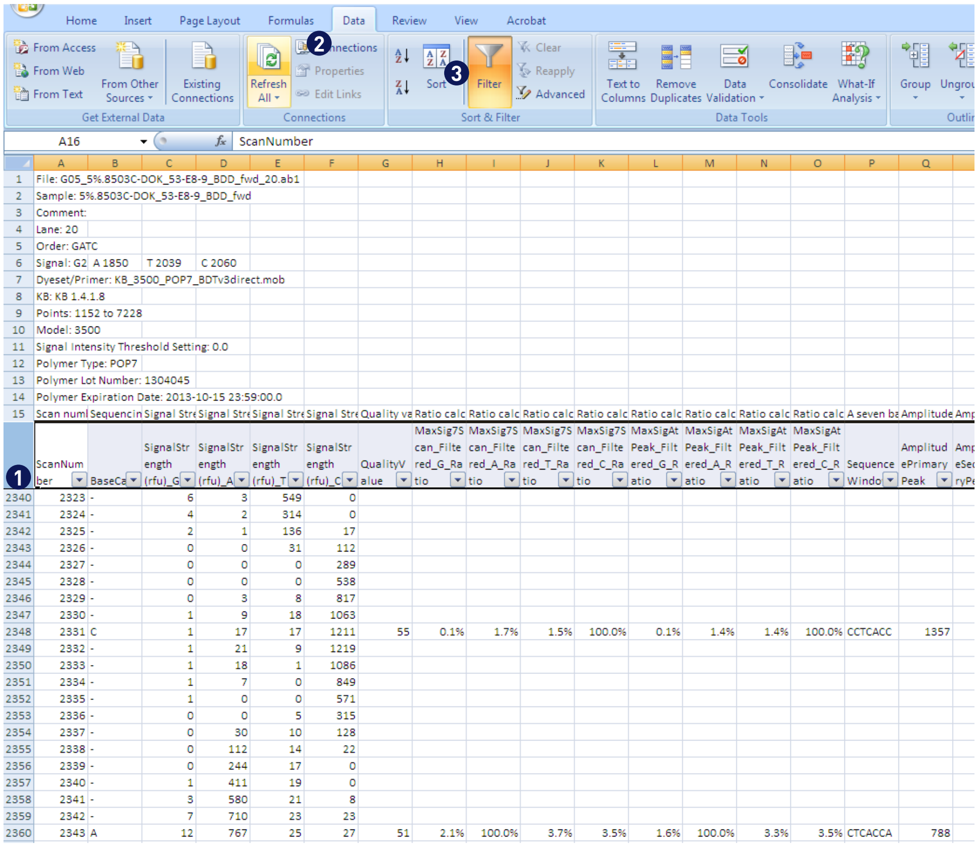
To facilitate identification of an allele (nucleotide) of interest (e.g., a SNP) in the table, the "Sequence Window" can be used to display each base in the center of a string of 7 nucleotides (Figure 15). Use the 7-base string as input in Excel's "Search" and "Find" functions. Use a * character if the base in the middle of the string is unknown or N,Y,R, etc. This string of 7 nucleotides can also be used in Sequence Analysis or Sequence



1 File: G05_5%.8503C-DOK_53-E8-9_BDD_fwd_20.ab1
2 Sample: 5%.8503C-DOK_53-E8-9_BDD_fwd
3 Comment:
4 Lane: 20
5 Order: GATC
6 Signal: G A 1850 T 2039 C 2060
7 Dyeset/Primer: KB_3500_POP7_BDTv3direct.mob
8 KB: KB 1.4.1.8
9 Points: 1152 to 7228
10 Model: 3500
11 Signal Intensity Threshold Setting: 0.0
12 Polymer Type: POP7
13 Polymer Lot Number: 1304045
14 Polymer Expiration Date: 2013-10-15 23:59:00.0
15 Scan nun Sequenci Signal St Signal St Signal St Signal St Quality v Ratio cal Ratio cal Ratio cal Ratio cal Ratio cal Ratio cal Ratio cal Ratio cal A seven l Amplituc Amplituc Base cal Base call for the se

	A	B	C	D	E	F	G
16	ScanNumber	BaseCall	SignalStrength (rfu)_(	SignalStrength (rfu)_/	SignalStrength (rfu)_1	SignalStrength (rfu)_(	QualityValue
340	2323	-	6	3	549	0	
341	2324	-	4	2	314	0	
342	2325	-	2	1	136	17	
343	2326	-	0	0	31	112	
344	2327	-	0	0	0	289	
345	2328	-	0	0	0	538	
346	2329	-	0	3	8	817	
347	2330	-	1	9	18	1063	
348	2331	C	1	17	17	1211	
349	2332	-	1	21	9	1219	
350	2333	-	1	18	1	1086	
351	2334	-	1	7	0	849	
352	2335	-	1	0	0	571	
353	2336	-	0	0	5	315	
354	2337	-	0	30	10	128	
355	2338	-	0	112	14	22	
356	2339	-	0	244	17	0	
357	2340	-	1	411	19	0	
358	2341	-	3	580	21	8	
359	2342	-	7	710	23	23	
360	2343	A	12	767	25	27	
361	2344	-	15	734	27	19	
362	2345	-	16	620	28	3	
363	2346	-	15	457	28	0	
364	2347	-	12	285	27	0	
365	2348	-	8	141	26	25	
366	2349	-	4	45	26	113	

Figure 12: Creating line plots of electropherograms. 1) Select cells A–F or B–F, 2) go to tab "Insert" and 3) select Line in the Chart section.



1 File: G05_5%.8503C-DOK_53-E8-9_BDD_fwd_20.ab1
2 Sample: 5%.8503C-DOK_53-E8-9_BDD_fwd
3 Comment:
4 Lane: 20
5 Order: GATC
6 Signal: G2 A 1850 T 2039 C 2060
7 Dyeset/Primer: KB_3500_POP7_BDTv3direct.mob
8 KB: KB 1.4.1.8
9 Points: 1152 to 7228
10 Model: 3500
11 Signal Intensity Threshold Setting: 0.0
12 Polymer Type: POP7
13 Polymer Lot Number: 1304045
14 Polymer Expiration Date: 2013-10-15 23:59:00.0
15 Scan numl Sequencin Signal Stre Signal Stre Signal Stre Signal Stre Quality va Ratio calc Ratio calc Ratio calc Ratio calc Ratio calc Ratio calc Ratio calc Ratio calc A seven ba Amplitude Amp

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
16	ScanNumber	BaseCa	SignalStrength (rfu)_G	SignalStrength (rfu)_A	SignalStrength (rfu)_T	SignalStrength (rfu)_C	QualityValue	MaxSig7Scan_Filtered_G_Ratio	MaxSig7Scan_Filtered_A_Ratio	MaxSig7Scan_Filtered_T_Ratio	MaxSig7Scan_Filtered_C_Ratio	MaxSigAtPeak_Filtered_G_Ratio	MaxSigAtPeak_Filtered_A_Ratio	MaxSigAtPeak_Filtered_T_Ratio	MaxSigAtPeak_Filtered_C_Ratio	SequenceWindo	AmplitudePrimaryPeak	AmpeSecryPe
2340	2323	-	6	3	549	0											
2341	2324	-	4	2	314	0											
2342	2325	-	2	1	136	17											
2343	2326	-	0	0	31	112											
2344	2327	-	0	0	0	289											
2345	2328	-	0	0	0	538											
2346	2329	-	0	3	8	817											
2347	2330	-	1	9	18	1063											
2348	2331	C	1	17	17	1211	55	0.1%	1.7%	1.5%	100.0%	0.1%	1.4%	1.4%	100.0%	CCTCACC	1357
2349	2332	-	1	21	9	1219											
2350	2333	-	1	18	1	1086											
2351	2334	-	1	7	0	849											
2352	2335	-	1	0	0	571											
2353	2336	-	0	0	5	315											
2354	2337	-	0	30	10	128											
2355	2338	-	0	112	14	22											
2356	2339	-	0	244	17	0											
2357	2340	-	1	411	19	0											
2358	2341	-	3	580	21	8											
2359	2342	-	7	710	23	23											
2360	2343	A	12	767	25	27	51	2.1%	100.0%	3.7%	3.5%	1.6%	100.0%	3.3%	3.5%	CTCACCA	788

Figure 13: Applying Filters to the data. 1) Click on row 16, 2) go to tab "Data" and 3) select Filter.

Scanner software to readily find a peak of interest.

Ratio of maximal signals in a 7-scan window

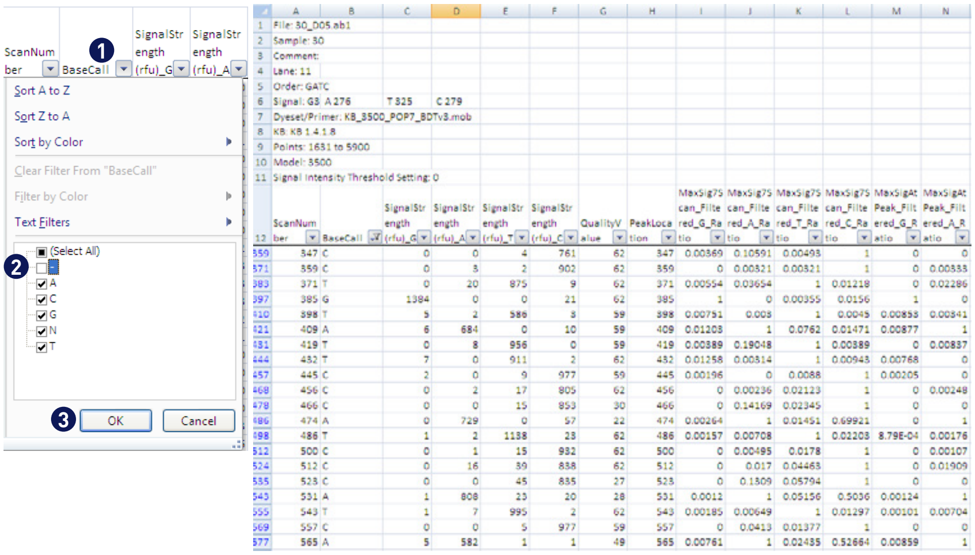
In columns I–L (Figure 15)the peak height ratios in a 7-scan window of the maximal signal between the primary (i.e., basecalled) peak and the maximal signal of either G, A, T, C respectively is shown. What exactly are these numbers?

The peak height ratio is calculated as the maximum of heights measured at the scan of the amplitude and the 3 scans upstream and downstream (hence 7-scan) of that particular location.

Figure 16 shows an example of the MaxSig7Scan Ratio calculation: at scan location 799 a peak was called "N"; the highest peak was an "A" trace with 555 RFU, followed by a "T" trace with 438 RFU and C (14 RFU) and G (12 RFU) in a 7-scan window (highlighted in yellow) which is centered at the peak location and extends for 3 more scans on either side (3+1+3 = 7 scans). The ratio is calculated by dividing the peak height of the "primary" (i.e., highest) peak by the basecalled or highest peak height of either peak trace (G in column I, A in column J, T in column K, and T in column L) in the 7-scan window. One caveat: in traces with sub-optimal spacing or mobility overlap between adjacent bases it is possible that the trailing or leading slope of a legitimate adjacent peak is considered in this calculation which may produce an artificially higher ratio.

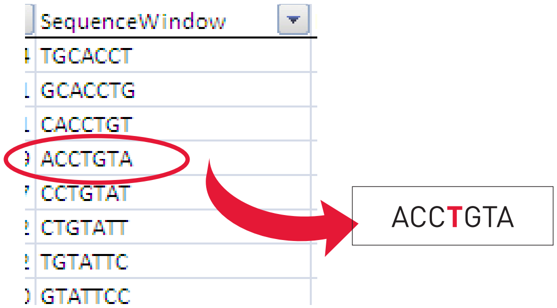
Ratio of signal in a 1-scan window at the basecall location

In columns M–P the peak height ratios in a 1-scan window of the maximal signal between the primary (i.e., basecalled) peak and the signal of either G, A, T, C, respectively, is shown (Figure 17).



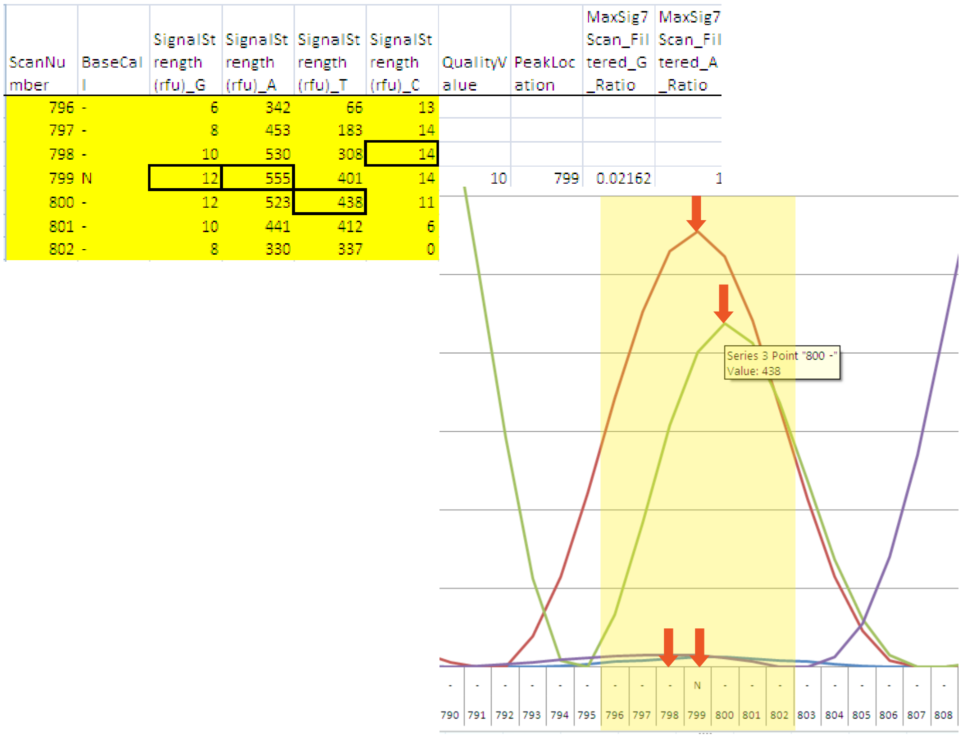
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	File: 30_D05.ab1													
2	Sample: 30													
3	Comment:													
4	Lane: 11													
5	Order: GATC													
6	Signal: G3	A 276	T 325	C 279										
7	Dyeset/Primer: KB_3500_POP7_BDTv3.mob													
8	KB: KB 1.4.1.8													
9	Points: 1631 to 5900													
10	Model: 3500													
11	Signal Intensity Threshold Setting: 0													
12	ScanNumber	BaseCall	SignalStrength (rfu)_G	SignalStrength (rfu)_A	SignalStrength (rfu)_T	SignalStrength (rfu)_C	QualityValue	PeakLocation	MaxSig7Scan_Filtered_G_Ratio	MaxSig7Scan_Filtered_A_Ratio	MaxSig7Scan_Filtered_T_Ratio	MaxSig7Scan_Filtered_C_Ratio	MaxSigAtPeak_Filtered_G_Ratio	MaxSigAtPeak_Filtered_A_Ratio
359	347	C	0	0	4	761	62	347	0.00369	0.10591	0.00493	1	0	0
371	359	C	0	3	2	902	62	359	0	0.00321	0.00321	1	0	0.00333
383	371	T	0	20	875	9	62	371	0.00554	0.03654	1	0.01218	0	0.02286
397	385	G	1384	0	0	21	62	385	1	0	0.00355	0.0156	1	0
410	398	T	5	2	586	3	59	398	0.00751	0.003	1	0.0045	0.00853	0.00341
421	409	A	6	684	0	10	59	409	0.01203	1	0.0762	0.01471	0.00877	1
431	419	T	0	8	956	0	59	419	0.00389	0.19048	1	0.00389	0	0.00837
444	432	T	7	0	911	2	62	432	0.01258	0.00314	1	0.00943	0.00768	0
457	445	C	2	0	9	977	59	445	0.00196	0	0.0088	1	0.00205	0
468	456	C	0	2	17	805	62	456	0	0.00236	0.02123	1	0	0.00248
478	466	C	0	0	15	853	30	466	0	0.14169	0.02345	1	0	0
486	474	A	0	729	0	57	22	474	0.00264	1	0.01451	0.69921	0	1
498	486	T	1	2	1138	23	62	486	0.00157	0.00708	1	0.02203	8.79E-04	0.00176
512	500	C	0	1	15	932	62	500	0	0.00495	0.0178	1	0	0.00107
524	512	C	0	16	39	838	62	512	0	0.017	0.04463	1	0	0.01909
535	523	C	0	0	45	835	27	523	0	0.1309	0.05794	1	0	0
543	531	A	1	808	23	20	28	531	0.0012	1	0.05156	0.5036	0.00124	1
555	543	T	1	7	995	2	62	543	0.00185	0.00649	1	0.01297	0.00101	0.00704
569	557	C	0	0	5	977	59	557	0	0.0413	0.01377	1	0	0
577	565	A	5	582	1	1	49	565	0.00761	1	0.02435	0.52664	0.00859	1

Figure 14: Condensing the data table to basecalled data points only.



SequenceWindow
TGCACCT
GCACCTG
CACCTGT
ACCTGTA
CCTGTAT
CTGTATT
TGTATTC
GTATTCC

Figure 15: The SequenceWindow lists 7-nucleotide strings, and can be used to facilitate finding a base of interest.



ScanNumber	BaseCall	SignalStrength (rfu)_G	SignalStrength (rfu)_A	SignalStrength (rfu)_T	SignalStrength (rfu)_C	QualityValue	PeakLocation	MaxSig7Scan_Filtered_G_Ratio	MaxSig7Scan_Filtered_A_Ratio
796	-	6	342	66	13				
797	-	8	453	183	14				
798	-	10	530	308	14				
799	N	12	555	401	14	10	799	0.02162	1
800	-	12	523	438	11				
801	-	10	441	412	6				
802	-	8	330	337	0				

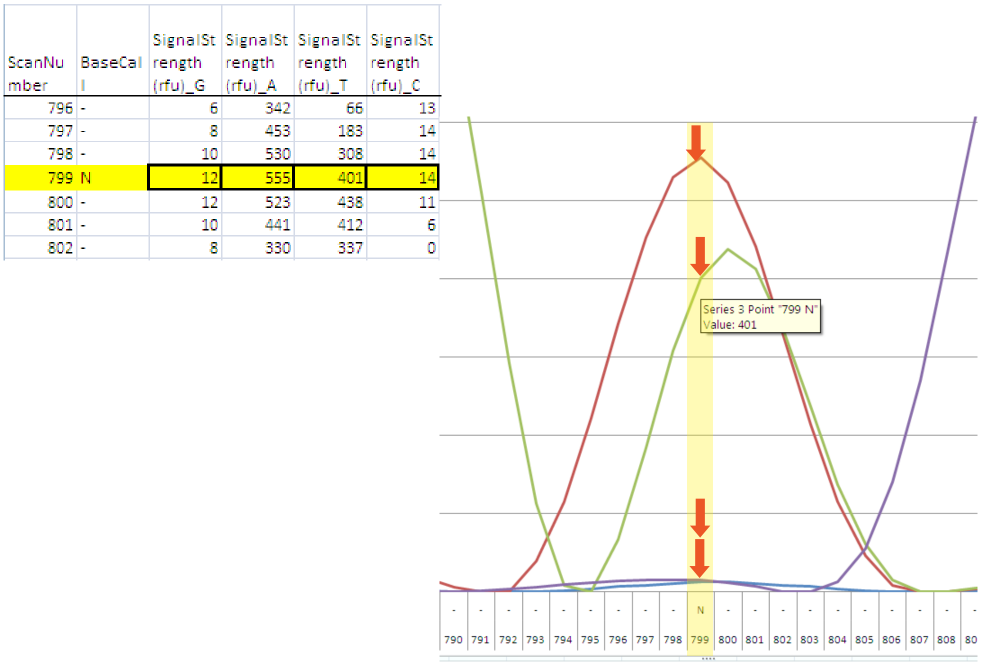
Figure 16: Ratios of maximal signal in a 7-scan window.

This ratio measurement is taken in a narrow window of one scan only. It may miss the amplitude of a peak under peak if it is outside this window. Therefore, both ratio measurements (7-scan and 1-scan) should be considered and possibly be combined (averaged), if necessary or warranted.

Data from the KB™ Basecaller (v1.4.1 and higher)

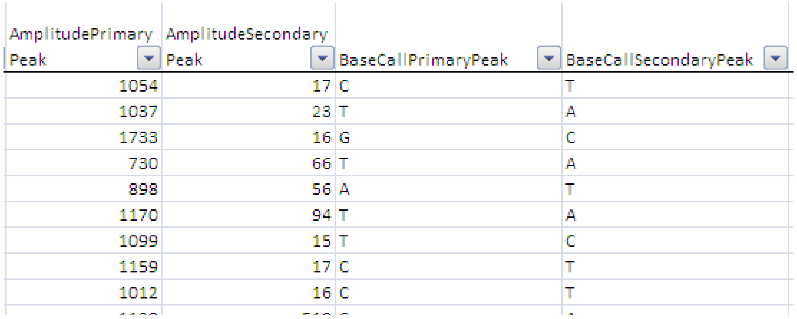
Columns Q, R, S, T are populated with the amplitude and sequence output data from the KB™ Basecaller (Figure 18). Note that the amplitudes of the primary and secondary peak may differ from the original signal strength (RFU) shown in columns C–F. This is due to the mobility and other noise correction of the trace data during the basecalling process.

Column G lists the Quality value (phd or phred score) of the basecall (Figure 19).



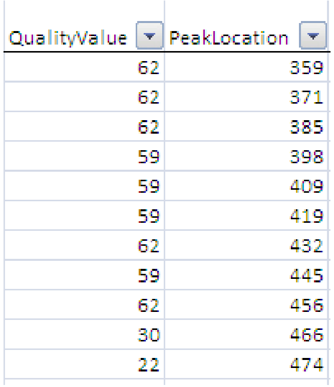
ScanNumber	BaseCall	SignalStrength (rfu)_G	SignalStrength (rfu)_A	SignalStrength (rfu)_T	SignalStrength (rfu)_C
796	-	6	342	66	13
797	-	8	453	183	14
798	-	10	530	308	14
799	N	12	555	401	14
800	-	12	523	438	11
801	-	10	441	412	6
802	-	8	330	337	0

Figure 17: The MaxSigAtPeak peak height ratio is determined by dividing the peak height of the primary peak (highest peak) by peak heights of either peak trace at the location (scan number) of the basecall.



AmplitudePrimaryPeak	AmplitudeSecondaryPeak	BaseCallPrimaryPeak	BaseCallSecondaryPeak
1054	17	C	T
1037	23	T	A
1733	16	G	C
730	66	T	A
898	56	A	T
1170	94	T	A
1099	15	T	C
1159	17	C	T
1012	16	C	T

Figure 18: Amplitudes and basecalls of primary and secondary peak as determined by KB™ Basecaller.



QualityValue	PeakLocation
62	359
62	371
62	385
59	398
59	409
59	419
62	432
59	445
62	456
30	466
22	474

Quality Values

The QV is a per-base estimate of the KB™ Basecaller accuracy.
The QVs are calibrated on a scale corresponding to:

$$QV = -10\log10(Pe)$$

where *Pe* is the probability of error.
The KB™ Basecaller generates QVs from 1 to 99.

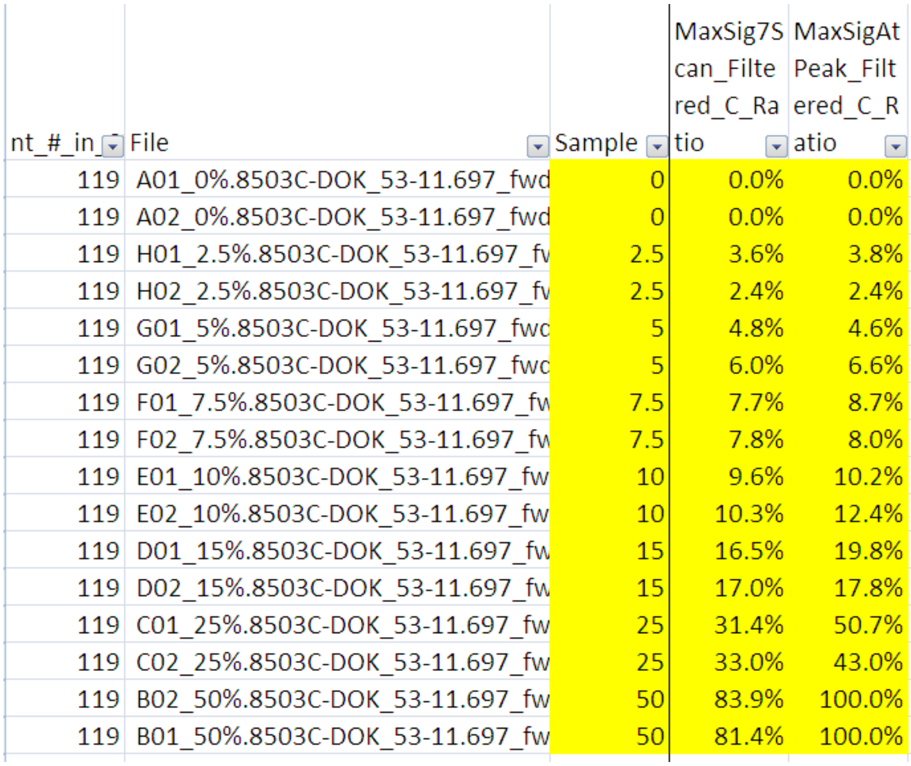
Quality Value	Probability the basecall is incorrect
10	10%
20	1%
30	0.1%
40	0.01%
50	0.001%

- Typical high-quality pure bases have QVs between 20–50
- Typical high-quality mixed bases have QVs between 10–20

Figure 19: The Quality values indicate the probability of an incorrect basecall of primary peak.

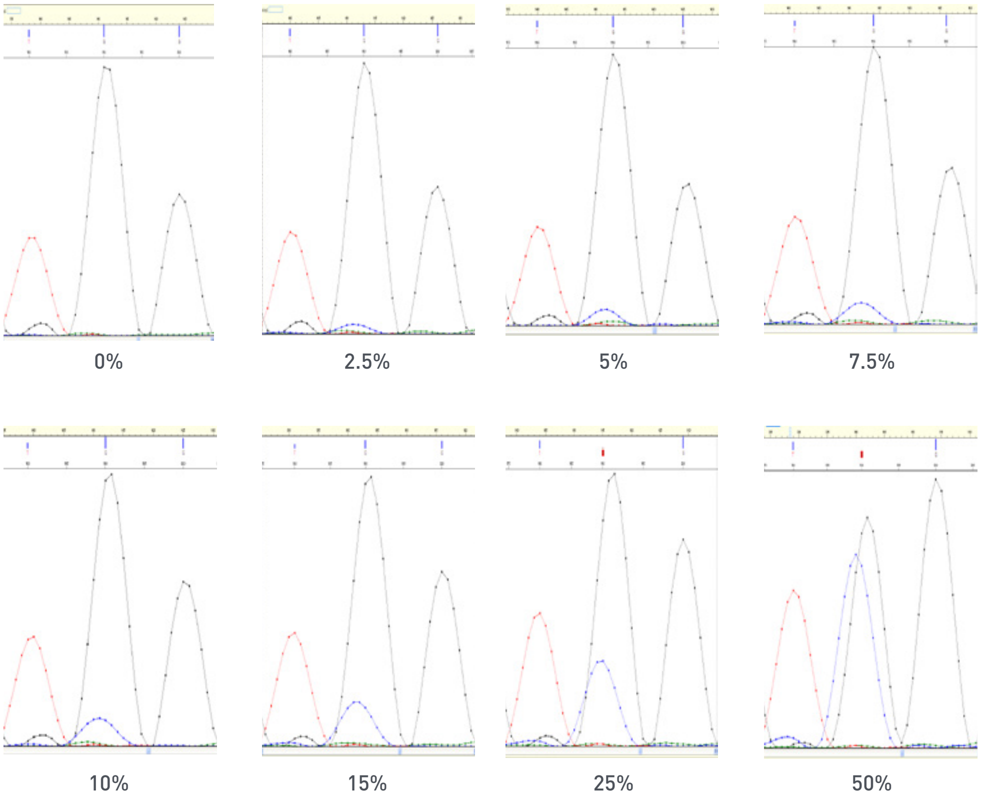
Measuring allele proportions by peak height ratios

To demonstrate the utility of the tool we have prepared genomic DNA mixtures of normal and mutant TP53 gene (exon 11) at various proportions and determined the peak height ratios between minor and major allele using the ab1PeakReporter tool. Figure 20 shows that in this particular allele situation the peak height ratios obtained from both channels (1-scan window or 7-scan window) correlated quite well up to 15%. A 5% level of mutant allele was clearly distinguishable from 0% (normal control; Figure 21).



nt_#_in	File	Sample	MaxSig7Scan_Filtered_C_Ratio	MaxSigAtPeak_Filtered_C_Ratio
119	A01_0%.8503C-DOK_53-11.697_fwd	0	0.0%	0.0%
119	A02_0%.8503C-DOK_53-11.697_fwd	0	0.0%	0.0%
119	H01_2.5%.8503C-DOK_53-11.697_fv	2.5	3.6%	3.8%
119	H02_2.5%.8503C-DOK_53-11.697_fv	2.5	2.4%	2.4%
119	G01_5%.8503C-DOK_53-11.697_fwc	5	4.8%	4.6%
119	G02_5%.8503C-DOK_53-11.697_fwc	5	6.0%	6.6%
119	F01_7.5%.8503C-DOK_53-11.697_fw	7.5	7.7%	8.7%
119	F02_7.5%.8503C-DOK_53-11.697_fw	7.5	7.8%	8.0%
119	E01_10%.8503C-DOK_53-11.697_fw	10	9.6%	10.2%
119	E02_10%.8503C-DOK_53-11.697_fw	10	10.3%	12.4%
119	D01_15%.8503C-DOK_53-11.697_fw	15	16.5%	19.8%
119	D02_15%.8503C-DOK_53-11.697_fw	15	17.0%	17.8%
119	C01_25%.8503C-DOK_53-11.697_fw	25	31.4%	50.7%
119	C02_25%.8503C-DOK_53-11.697_fw	25	33.0%	43.0%
119	B02_50%.8503C-DOK_53-11.697_fw	50	83.9%	100.0%
119	B01_50%.8503C-DOK_53-11.697_fw	50	81.4%	100.0%

Figure 20: Peak height ratios.



0%

2.5%

5%

7.5%

10%

15%

25%

50%

Figure 21: Sequencing electropherograms of DNA mixtures prepared at various ratios of wt and mutant allele "697" in exon 11 of the human p53 gene as viewed in Sequence Scanner software. Note that the mutant allele was "called" as "S" at the 25% and 50% level but not below these ratios using the KB™ Basecaller.

	QSVanalyzer	ab1PeakReporter
Number of alleles	Limited to predefined positions	All bases in trace file
Number of .ab1 files that can be analyzed	Multiple (maximum # not specified) QSV analysis requires presence of homozygous controls for either variant	96 (maximum upload per processing)
Table of peak height data of primary and secondary peaks	Yes (see Figure 6, columns B and C)	Yes (requires that .ab1 file is analyzed with KB™ Basecaller v1.4)
Compatible data files	.ab1, .scf	.ab1
Peak traces displayed	Yes, in comprehensive HTML report and in separate window as .png file	No, but can be created as a line graph in Excel using .csv file with complete data points
Output reports	Folder with trace data, comprehensive QSV report in HTML and table (.xls) with raw and normalized peak heights and ratios	Zip folder with .csv file that opens in Microsoft Excel®
Suitable for quantitative assessment of SNPs, paralogous variant analysis and copy number variants	Yes (see Carr et al. 2009 for details)	Delivers peak height data and peak height ratios for customized downstream analysis
Suitable for methylation analysis (sequencing of bisulfite-converted DNA)?	Can potentially provide allele ratios CpG to TpG (UpG). Delivers normalized peak height data for customized downstream analysis	Delivers peak height data for customized downstream analysis
Suitable for minor allele quantification (somatic or emerging mutations)?	Possible when used with appropriate calibrator controls, replicates and customized data analysis (polynomial regression), see Figure 7	Delivers peak height data for customized downstream analysis

Table 1: Summary of features available in the QSVanalyser application and the ab1PeakReporter tool.

Conclusions

This application note shows tools and methods for extracting and using peak height data from fluorescent Sanger sequencing traces for determination of allele ratios or allele quantification. Table 1 summarizes the features of the two software applications presented.

References

[1] Carr IM*, Robinson JI, Dimitriou R et al. (2009) *Bioinformatics*, 25 (24):3244–3250. http://bioinformatics.oxfordjournals.org/content/25/24/3244.long

[2] White paper: Applied Biosystems Genetic Analysis Data File Format http://www6.appliedbiosystems.com/support/software_community/ABIF_File_Format.pdf

Find out more at **lifetechnologies.com**

life technologies™