

# **DNA Sequencing Research Group (DSRG)**

## **Single Nucleotide Polymorphism (SNP) Study 2002 Results**

**D. Leviten<sup>1</sup>, J. Hawes<sup>2</sup>, T. Hunter<sup>3</sup>,  
E. Jackson-Machelski<sup>4</sup>, K. L. Knudtson<sup>5</sup>,  
R. Pershad<sup>6</sup>, D. Spicer<sup>7</sup>, D. Bartley<sup>8</sup>,  
G. Grills<sup>9</sup>, M. Robertson<sup>10</sup>,  
T. Thannhauser<sup>11</sup>**

**<sup>1</sup>ICOS Corp., <sup>2</sup>Indiana University, <sup>3</sup>Vermont Cancer  
Center, <sup>4</sup>Washington University School of  
Medicine, <sup>5</sup>University of Iowa, <sup>6</sup>UT MD Anderson  
Cancer Center, <sup>7</sup>Prolinx Inc., <sup>8</sup>John Hopkins  
University, <sup>9</sup>Harvard University, <sup>10</sup>University of  
Utah, <sup>11</sup>Cornell University**

# ABSTRACT

Single nucleotide polymorphisms (SNPs) are new to the sequencing core environment and are fast becoming an important tool for researchers in both academic and commercial settings. The DSRG was particularly interested in and concerned about how DNA sequencing core facilities would be able to perform SNP sequencing projects. The goal of this study, therefore, was to determine whether SNPs and different ratios of SNP mixtures could be sequenced accurately using the equipment and chemistries currently being used in participating ABRF member laboratories. Some major questions to be addressed were: 1) how reliably are SNP's 'called' and 2) with what sensitivity can SNPs be correctly detected in a core setting. Initially, the DSRG launched an internal study to determine whether our current equipment and chemistries could accomplish the above goal. It was determined that the goal could be accomplished, and these results were given at the ABRF 2001 conference. Since then, an external study has been launched, and data from this study were compared and contrasted to the results from the internal study.

# INTRODUCTION

**SNP (single nucleotide polymorphism) detection, especially ratio calling, is fairly new to the sequencing core environment. Over the last year, new technologies, software and products have been introduced and more will follow. To successfully complete the goal of this study, the following questions must be answered: (1) Can laboratories detect SNPs and SNP mixtures? (2) Can the SNP mixtures be detected accurately with a single pass of sequencing? (3) Is there a difference between slab gel versus capillary sequencing in evaluating and correctly calling the SNPs? (4) Are there differences in the various chemistries in evaluating and correctly calling the SNPs? (5) What is the lowest percent mixture of SNPs that is accurately detectable in the current sequencing systems available? (6) What software programs are available to analyze SNP data? Can these programs also determine the ratios of the mixtures? Are the programs easy to use and interpret? This study is a snapshot of what is currently being used for SNP ratio calling by core laboratories. By answering the above questions thoroughly, our goal is to highlight methods of choice for protocols, software, chemistries, and machine platforms for calling SNP ratios.**

# METHODS

The study utilized 2 sets of SNP clones, 4A & 6A and 4M & 5M. Each set contained the wildtype (homozygous) for one sequence mixed at some ratio with the wildtype (homozygous) sequence. The ratios were not revealed, but were left as a challenge to identify. However, the sequences for both the wildtypes were included with the samples. Set #1 (4A&6A) contained one SNP and Set #2 (4M&5M) contained 4 SNPs. Details about the sequences, the location of the SNPs and the base composition of the SNPs are listed below. For more information, see the ABRF website ([www.abrf.org](http://www.abrf.org)).

There were 7 ratios for each set. Universal primers were used for sequencing: M13 forward (-20) (TGTAACGACGGCCAGT) and M13 reverse (AACAGCTATGACCATG).

The samples were PCR products of 130 bp (base 34 is the SNP of G/A) and 310 bp (SNPs at base 40 A/G, 217 C/T, 272 T/C and 307 C/G).

These maize clones (plasmids) were generously donated by Dr. Steve Kresovich from the Institute of Genomic Diversity at Cornell University. The first set contained one SNP, the second contained 4 SNPs. These clones were selected for growth and plasmid were purified using Qiagen's Midi Kit. Plasmid DNA was dissolved in Qiagen Elution Buffer (10 mM Tris-HCl). 100 ng of DNA isolated from the plasmid preps were PCR amplified using 10  $\mu$ l of M13(-21) and M13 reverse primers (100  $\mu$ g/ml), 10  $\mu$ l 10X dNTP mix, 10  $\mu$ l 10X PCR buffer stock, and 1  $\mu$ l of Taq Polymerase (5 U/ $\mu$ l) in a 100  $\mu$ l reaction volume. The thermal cycling protocol was: 94 C for 4 min. followed by 30 cycles of 94 C for 30 sec., 65 C for 30 sec and 72 C for 2 min. PCR samples were resolved on a 2% agarose gel and then purified utilizing the QIAquick PCR Purification Kit (Qiagen #28704).

The amplicons were quantified using a UV spectrophotometer and concentrations were adjusted to 5 ng/ $\mu$ l for Set #1 and 15 ng/ $\mu$ l for Set #2. This concentration was based on a general rule of thumb of ~10 ng per 100 bases of PCR product per sequencing reaction. Mixtures were made in the following ratios:

Clone 4A or 4M	Clone 6A or 5M	Tube Nomenclature
100	0	A or J
80	20	B or K
70	30	C or L
60	40	D or M
50	50	E or N
40	60	F or O
30	70	G or P
20	80	H or Q
0	100	I or R

**Note:** Set #1 (4A&6A) is in **Yellow**. Set #2 (4M&5M) is in **Red**.

7  $\mu$ l of each ratio mixture was aliquoted into individual tubes that were labeled 4AA through 4AI and 4MJ through 4MR, for a total of 18 tubes. Additional sample was sent to individuals trying more than one chemistry or instrument.

For the external study, the study was announced on the ABRF e-mail listserver and a similar listserver for sequencers in the UK and was also posted on the ABRF website. Once a sample was requested, all survey and data information remained anonymous using a unique identifier chosen by the participating laboratory.

# RESULTS & DISCUSSION

Prior to opening the study to the ABRF membership, an internal study was first conducted by the members of the DSRG to assess potential challenge areas in conducting this study. SNP mixtures were sequenced by DSRG members using the following equipment and chemistries: ABI 373, ABI 377, ABI310, ABI3700, Big Dye Terminator v2 and Big Dye Primer. Analysis methods included visual inspection, Sequencher software (GeneCodes) and Polyphred software (Phil Green/David Gorton, University of Washington). Given that the location of the SNP/SNP mixture would be disclosed to participants in the study, visual analysis was fine for determining that the SNP was detected. The drawback to visual analysis is that there is no way to quantify the ratios of the mixtures (areas of the peaks on the chromatograms).

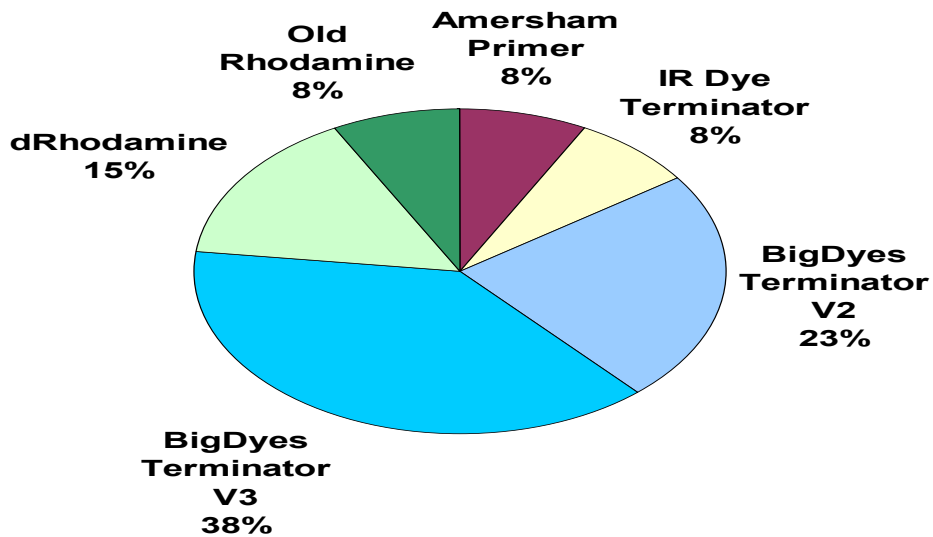
Sequencher and Polyphred were tested for their abilities to detect SNPs and quantify the ratios. They were also evaluated on their ability to enable conclusive judgments on the quality and ratio of the ABI sequencing software automated call of the SNP. Sequencher was able to call the secondary peaks based upon thresholds manually set by the user. After extensive testing, we concluded that it was too difficult to 'train' Polyphred on the sample set we had generated. Mixtures could be sequenced with current chemistries and equipment. Both strands needed to be sequenced for complete confidence on the calls. Thus, it

**was determined that the challenge area for the open study would be the choice of software used to analyze the SNPs and it was decided that Sequencer would be used.**

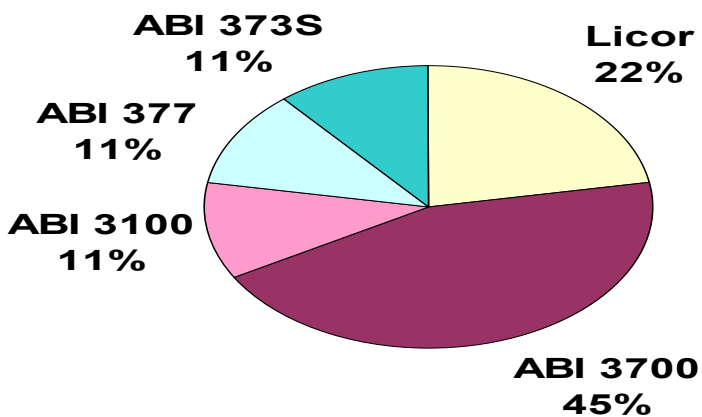
**The results of the internal study, therefore, showed that there were not significant differences in results between slab gel and capillary sequencing or between Big Dye Terminator and Big Dye Primer sequencing. The lowest ratio that bordered giving confident calls was 70:30. Sequencer software was the most accurate for calling ratios between the three analysis options tested.**

**The open study was launched in September 2001, and closed February 2002. Thirty-one individual laboratories requested sample sets, four of these laboratories indicated use of more than one instrument platform. Twelve of these laboratories submitted surveys, however, not all surveys were accompanied by chromatogram data. A total of 373 chromatograms (or equivalent, depending on instrument) were submitted.**

**Most participants (61%) used Big Dye Terminator chemistries (Figure 1) and 78% of the submissions were generated using ABI instruments (Figure 2). Thus, it appears that most laboratories that are conducting SNP detection with their sequencers are using Applied Biosystems instruments and chemistries.**



**Figure 1. DNA sequencing chemistries used by the participants of the study.**



**Figure 2. DNA sequencing instruments used by the participants of the study.**

A majority of the labs used visual inspection alone or visual inspection with a software component (Figure 3). The most popular software used for SNP calling was Sequencher. Participating laboratories were able to call Set #1, which consists of 1 G/A SNP (Figure 4), more accurately than Set #2, which consists of 4 SNPS (Figure 5). The machines that called the most correct ratios were the ABI 373S with old rhodamine chemistry (100% correct calls for 4A:6A ratios) and a Li-Cor using only visual inspection (8/9 correct calls in 4A:6A ratios, and 1/9 correct calls in 4M:5M ratios). SNP mixtures sequenced in both directions helped resolve problems unique to particular chemistries (i.e., smaller G peak after A peak). The bidirectional approach also helped resolve template specific problems such as sequence specificity that may influence the peak morphology and therefore influence the SNP ratio call.

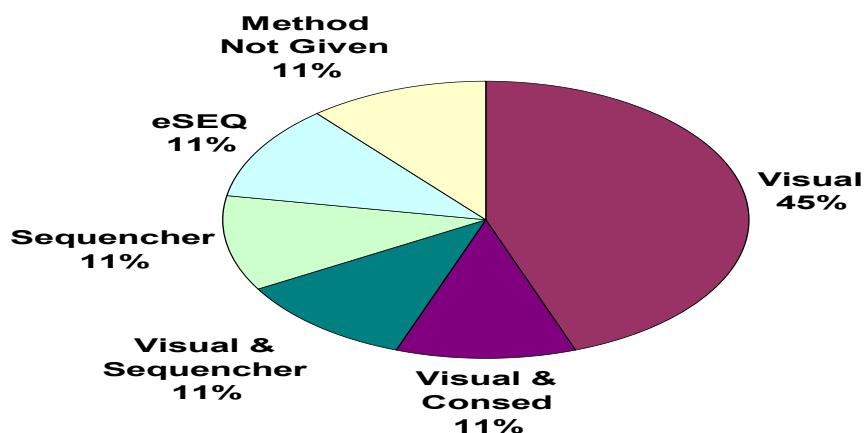


Figure 3. SNP calling method used by the participants of the study.

**There were two laboratory participants that called the ratios most accurately and were significantly better at calling the SNP ratios than the rest of the participants. The 373S with old Rhodamine dye chemistry (both from Applied Biosystems) called all of the Set #1 ratios correctly (4A:6A). Unfortunately, this lab did not indicate what method they used for analyzing their ratios and did not enter estimates for Set #2; therefore this lab made 9 estimates rather than the 18 estimates made by all the other labs. The other lab that called the most correct ratios (9/18 total, 8/9 from Set #1 and 1/9 from Set #2) determined the ratios using visual inspection alone. These two labs used very different equipment and dye chemistries. We would need more details about their methods to pinpoint what increased the probability of success for these two laboratories. A key factor might be an experienced eye for calling SNPs.**

**Differences in slab gel versus capillary may also be dependent upon the experience of the laboratory. The results show that in some instances gels may work better than capillaries or vice a versa (see Figures 4& 5). Most labs had difficulties calling the correct mixture percentages. Most correct calls were for mixtures close to 50:50. Calls became more inaccurate at a 70:30 mixture and below.**

Accuracy by Machine Type, Chemistry and Analysis Method

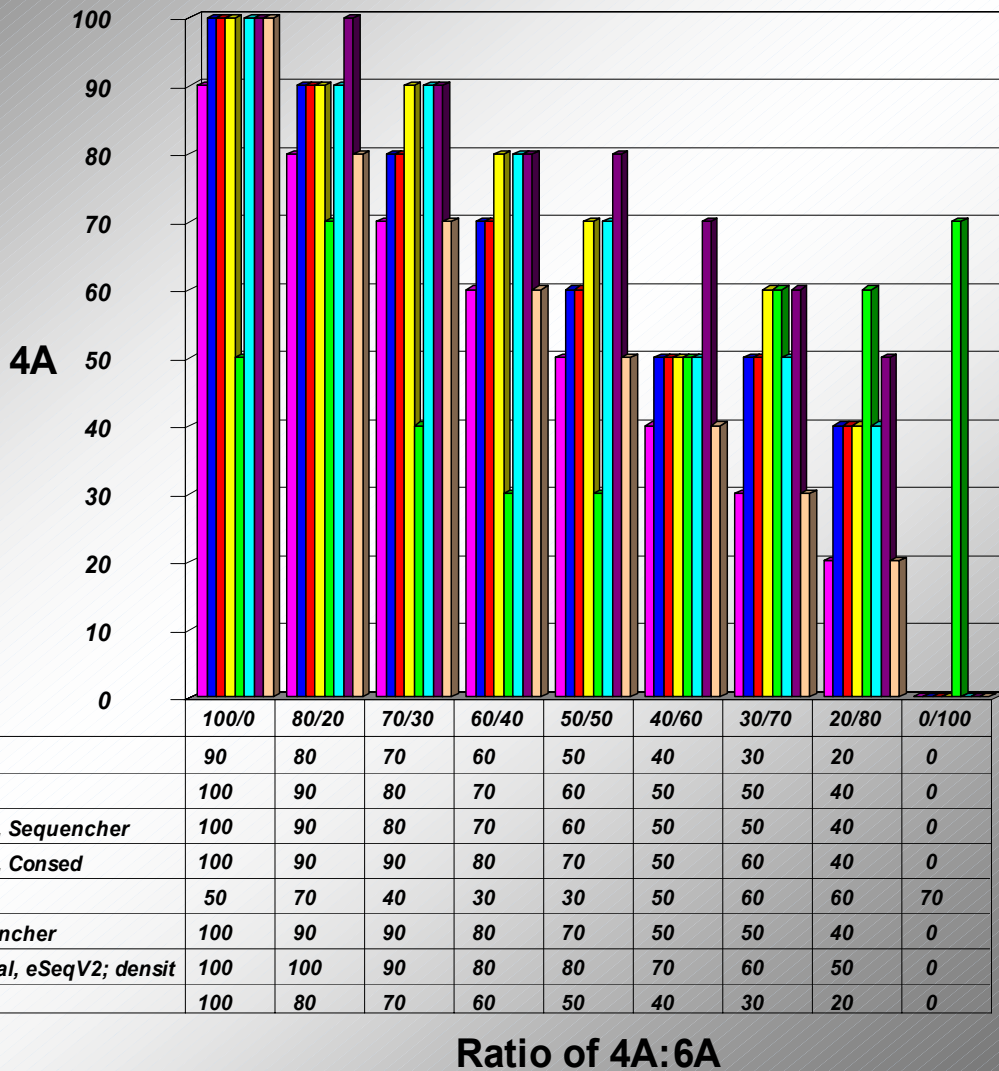
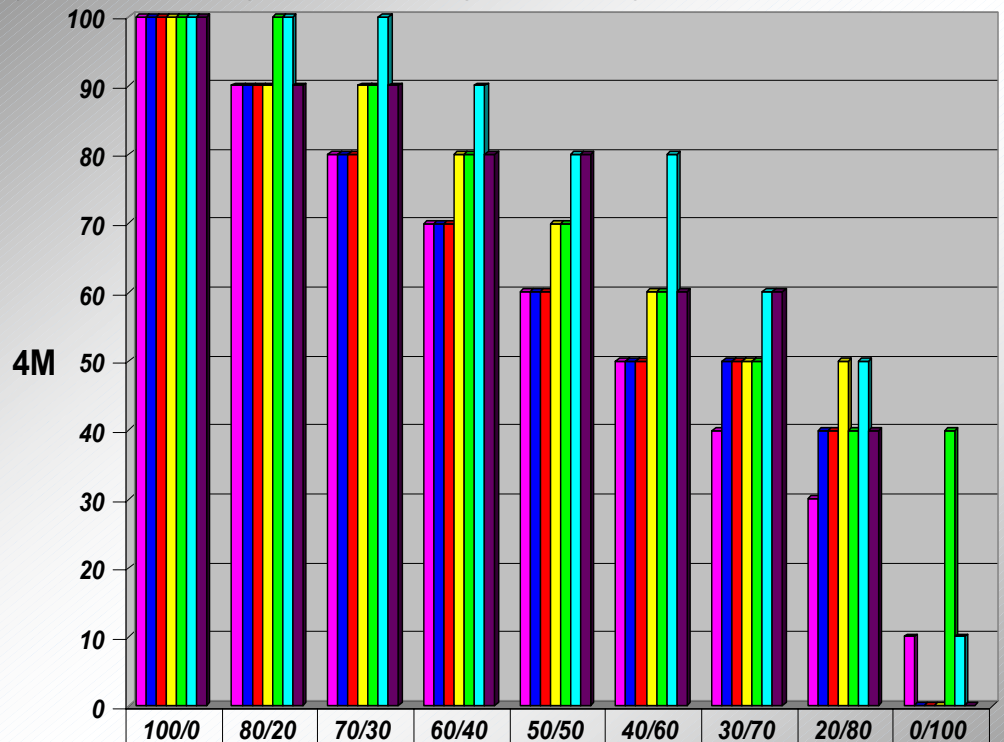


Figure 4. Comparison of the accuracy of different instruments and chemistries in predicting mixed ratios of a G/A polymorphism.

### Accuracy of Machine Type, Chemistry and Analysis Method



	100/0	80/20	70/30	60/40	50/50	40/60	30/70	20/80	0/100
Li-Cor-Visual	100	90	80	70	60	50	40	30	10
3700-BD3-Visual	100	90	80	70	60	50	50	40	0
3700-BD3-Visual, Sequencher	100	90	80	70	60	50	50	40	0
3700-BD2-Visual, Consed	100	90	90	80	70	60	50	50	0
3100-BD3-Visual	100	100	90	80	70	60	50	40	40
3700-BD3-Sequencher	100	100	100	90	80	80	60	50	10
Li-Cor 4200-Visual, eSeqV2; densit	100	90	90	80	80	60	60	40	0

Ratio 4M:6M

**Figure 5. Comparison of the accuracy of different instruments and chemistries in predicting mixed ratios of a mixture of 4 polymorphisms (A/G, C/T, T/C and C/G).**

**One laboratory submitted 176 chromatograms that were generated by using three different chemistries on two different sequencing platforms. This subset of data, therefore, was used to examine the relative effectiveness of these chemistries and platforms in the detection of SNPs. The data was interrogated using Sequencher software using the “call secondary peaks” function that assigns an ambiguity code to mixed base peaks on the basis of the percentage of secondary peak height to the main peak height.**

**Each chromatogram was analyzed using the call secondary peak feature starting at 100% and dropping the percentage by varying degrees until an ambiguity code (IUPAC-IUB) was assigned. The percentage at which the ambiguity code was assigned was noted. For a true heterozygous mixed base, one would expect the call to be assigned at 100% if the peak height of both peaks was equivalent. For a SNP that is mixed at 80:20 ratio, Sequencher assumes the main peak is 1.00 and compares the percentage of the secondary peak to the main peak. Therefore the expected percentage at which the lower ratio would be called should be 25% ( $1.00 \times 20/80$ ).**

**As shown in Figure 6, the C/G SNPs were generally more accurately called from data generated using the ABI Prism Model 377. Conversely, the A/G SNPs were more accurately called when the ABI Prism Model 3700 was used to generate the data. The use of Big Dye v3 seemed to give better sensitivity than either the Big Dye v2 or dRhodamine in SNP detection when used on either the slab or capillary gel platform.**

The results shown in Figure 6, however, were difficult to interpret because there are many instances where the peak heights did not reflect the true heterozygosity of the allele or the SNP ratio in this case. Different sequencing chemistries can result in varying peak heights in all four bases across the chromatogram. Dye primer chemistry usually gives more even peak heights than dye terminator chemistry. The many different versions of dye terminator chemistry can also produce chromatograms of varying peak heights. Thus, the measured SNP ratios rarely matched the expected.

In addition, this data set was unbalanced in that results were not submitted for all sequencing chemistry and sequencing platform combinations for each SNP ratio. Thus, a definitive assessment of chemistry and sequencing platform could not be conducted.

Figure 7 suggests that the LiCor and ABI 373s instruments were better than the other instruments in giving data that permit accurate calling of SNPs. This is unlikely, however, because these data represent a small number of samples provided by a single investigator. A more likely explanation is that these two investigators possessed experience in calling SNPs. Most of the submissions were run on ABI 3700 instruments using a variety of sequencing chemistries and calling methods. Given the small number of submissions for each instrument/ chemistry/calling method combination, it was not possible to demonstrate that any combination was significantly better than another in accurately detecting SNPs.

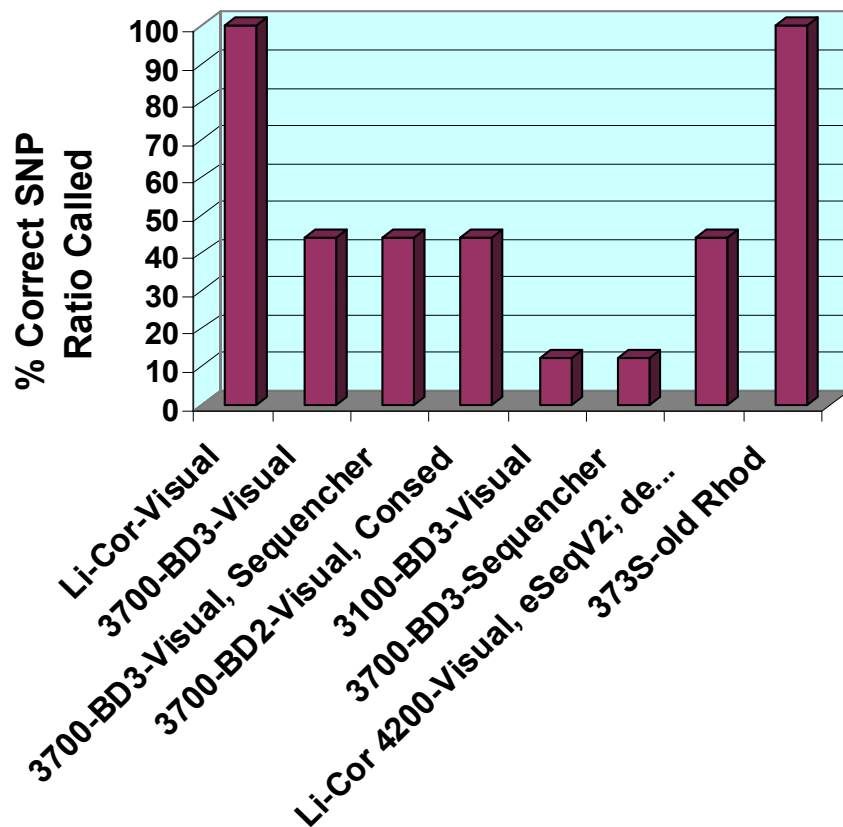
		0052 A/G MUTATION						
Sample ID		4Ab	4Ac	4Ad	4Ae	4Af	4Ag	4Ah
Ratio		80G:20A	70G:30A	60G:40A	50A:50G	40A:60G	30A:70G	20A:80G
Minor Peak		25%	43%	67%	100%	67%	43%	25%
		14%	22%	40%	66%	98%	96%	62%
		13%	21%	32%	48%	80%	94%	60%
		13%	19%	30%	42%	70%	94%	60%
		8%	16%	26%	40%	62%	86%	56%
		5%	10%	21%	34%	58%	86%	54%
		2%	5%	19%	32%	54%	80%	52%
							72%	38%
							60%	34%
								30%

		0052 C/G MUTATION						
Sample ID		4Ak	4Al	4Am	4An	4Ao	4Ap	4Aq
Ratio		80C:20G	70C:30G	60C:40G	50C:50G	40C:60G	30C:70G	20C:80G
Minor Peak		25%	43%	67%	100%	67%	43%	25%
		60%	62%	92%	96%	94%	72%	40%
		50%	60%	86%	94%	90%	56%	32%
		32%	46%	80%	88%	76%	50%	30%
		32%	44%	68%	78%	74%	40%	28%
		30%	34%	64%	66%	70%	30%	
		14%		40%	50%			
		<14%						

COLOR KEY
v3 FORWARD 3700
v3 REVERSE 3700
v2 FORWARD 3700
v2 REVERSE 3700
dRhod FORWARD 3700
dRhod REVERSE 3700
v3 FORWARD_377
v3 REVERSE_377
dRhod FORWARD_377
dRhod REVERSE_377

**Figure 6. Percentages of secondary peak calls used to predict an A to G (top) or C to G (bottom) SNP from data (176 chromatograms) submitted by one laboratory generated using three different chemistries and two different sequencing platforms. Minor Peak=Expected Percentage of the Minor Peak.**

These data were interrogated using the “call secondary peaks” function in Sequencher. If this function assumes the main peak is 1.00 and compares the percentage of the minor peak to the main peak, then the expected percentage at which the minor peak would be called for the 80/20 ratio would be 25% (1.00x20/80).



**Figure 7. Comparison of total number of accurate SNP ratio calls by different machine types, dye chemistries and analysis methods.**

# CONCLUSIONS

The number of samples submitted for this study was small, but provide a snapshot of how facilities are addressing the needs to sequence SNPs. In this study most laboratories were able to correctly identify heterozygotes by DNA sequencing.

The choice of instrumentation did not seem to exert any significant influence on the ability to identify the SNP ratio correctly. Some of the SNPs were called more consistently on the capillary platform while other SNPs were more accurately called from slab gel data. There were also no underlying trends that could be determined by looking at accurate calls versus sequencing chemistry.

Most laboratories had difficulty in calling the SNP ratios correctly and visual inspection was the primary method for identifying the percentage of the minor SNP. In this study Sequencher was the preferred software for calling SNPs and the most correct calls were for the 50:50 mixtures. In most cases the confidence in SNP ratios calls was not reliable below the 70:30 mixture. It is recommended that sequencing should be performed in both directions to make confident SNP calls.