

## **Mutation Detection from DNA Sequence Traces with Mutation Surveyor and Mutation Explorer Software**

ChangSheng (Jonathan Liu) and Shouyong Ni

SoftGenetics, 200 Innovation Blvd, Suite 241, State College, PA 16803, USA.

Email: [liu@softgenetics.com](mailto:liu@softgenetics.com), phone 1-814-237-9340, Web [www.softgenetics.com](http://www.softgenetics.com)

Mutation detection is increasingly undertaken as a tool for a wide spectrum of research especially in cancer diseases<sup>1,2,3,4</sup>, disease association and clinical diagnostics. The pharmaceutical industry spends billions of dollars to locate the mutated genes associated with particular diseases. There are many indirect methods, such as dHPLC, to determine mutation existence in a DNA sample, but these indirect methods do not provide all the necessary context information required.

DNA sequencing, a direct method, is a fundamental technology that not only detects all mutation types, but also provides critical mutation location in the sample DNA sequence. Unfortunately the use of sequence data and the comparison of patient/sample sequence to wild type or normal sequence traces have been cumbersome and difficult, limiting the use of this primary technology in mutation detection. DNA variant detection has been limited to analytics that were developed as assembly tools for the human genome project. Typically these tools are based upon comparison of the text base call, which at Phred 20 does not have the required accuracy for Mutation detection. At Phred 20 the base calling error rate is 1%, while hereditary mutations have an average occurrence of .08%, somatic mutation perhaps 1 in a million bases. Therefore the use of these analysis tools also require manual visual inspection, which is too time consuming and inaccurate for the pace of today's research requirements.

Direct sequence trace comparison methods have been discussed in a few scientific papers have discussed the use of sequence traces to locate heterozygote and homozygote point mutations, however there is no known paper discussing detection of insertion and deletion mutations, especially heterozygous insertions and deletions, via direct trace comparison.

The following are examples of the many indirect methods available to detect sequence variation in a specific region of DNA from multiple samples. One such series of indirect methods is referred to as mutation discovery methods. Mutation discovery methods detect the relative peak shifting when a mutation sample is compared to wide-type reference DNA. Mutation discovery methods include denaturing gradient gel electrophoresis (DGGE), denaturing high performance liquid chromatography (DHPLC), temperature gradient capillary electrophoresis (TGCE), heteroduplex analysis (HD), the analysis of single stranded DNA conformation polymorphism (SSCP), and chemical or enzyme cleavage of the mismatch (CECM). However, these mutation discovery methods are not able to provide the specific mutation location in the DNA sequencing. Therefore, the mutation discovery methods cannot tell where the mutation has taken place or the type of mutation. Because these discovery methods are indirect they typically require mutation confirmation by DNA sequencing.

Another series of indirect methods is referred to as mutation genotyping. An example of mutation genotyping is the single base extension method, which detects mutation type when the DNA sequence is known. The above two series of indirect methods involve comparing two peaks of an electropherogram.

The direct method of Mutation detection is DNA sequencing, which can provide definitive mutation location and type of mutation in the sample versus wild type or normal.

Thus far, mutation discovery utilizing DNA sequencing has involved a large amount of calculation and extensive data manipulation, and has been fraught with inaccuracies and tedium. A software program for automatic detection of DNA variants from sequence trace data appears to be the most prudent and efficient method for discovery of disease causing genetic variants. There are academic software packages for mutation detection using trace data such as PolyPhred<sup>5</sup>, an analytical software package that utilizes the Phred score to detect heterozygous point mutations; Trace-diff<sup>6</sup> in the Staden software package which subtracts the reference traces from the sample traces to indicate possible variants. However, these academic software programs detect only a specific mutation type (e.g. homozygous and heterozygous point mutations), typically utilizing a specific chemistry. None of them are capable of detecting all kinds of

mutations with all chemistries. Other drawbacks to the available academic software programs are lack of flexibility, high learning curve, the requirement of visual inspection of final results due to relatively high error rate and difficulty of use.

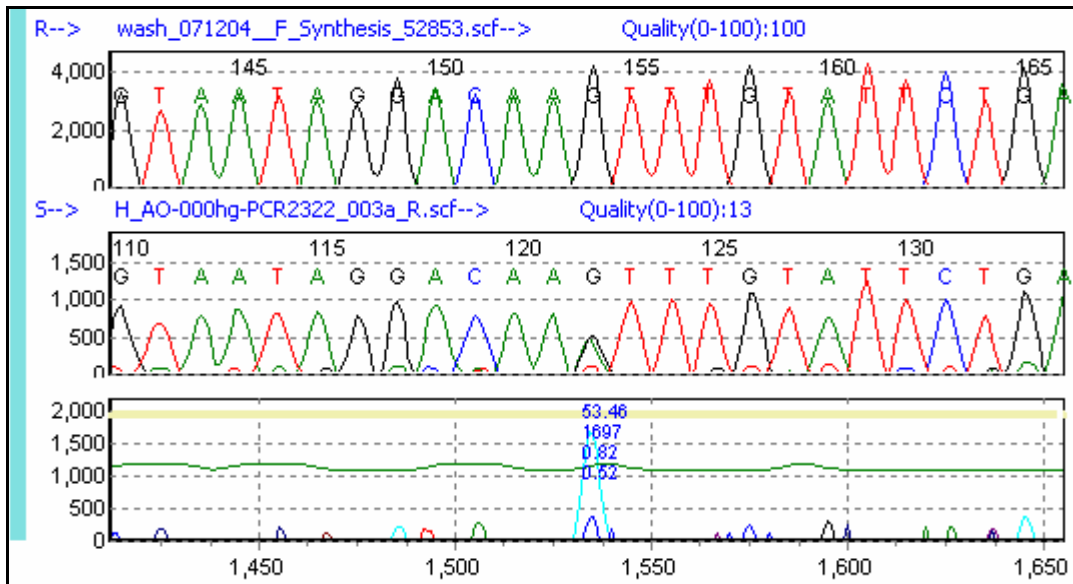
There are a few commercially available software packages for the detection of mutations from DNA sequence traces. SoftGenetics is a pioneer in mutation detection from DNA sequence traces. In collaboration with the Vogelstein/Kinzler laboratory of the Johns Hopkins School of Medicine, SoftGenetics has developed programs for DNA Variant analysis directly from sequence traces. The SoftGenetics software, Mutation Surveyor and Mutation Explorer, are currently widely used in the academic research<sup>1-4, 7,8,9,10,11,12</sup> and clinical diagnostics. Both SoftGenetics' programs provide high accuracy, high sensitivity, great speed, and ease of use. The software detects not only homozygous and heterozygous point mutations but also homozygous and heterozygous indels. In addition to Mutation Surveyor/Explorer software, SeqScape<sup>13</sup> from Applied Biosystems and Agent<sup>14</sup> from Paracel Inc. are software packages for mutation detection.

We will now discuss the technical details of the Mutation Surveyor software in this paper. Both Mutation Surveyor and Explorer software programs perform a physical comparison of the sequence trace of the patient/sample to a known reference trace, using several key factors to locate point mutations. Migration time changes of the sample versus the reference are used for homozygous and heterozygous indel detection. The program provides a log scale confidence score for each found variant that is based upon statistical theory used in electrical signal processing.

Mutation Surveyor and Explorer are identical in functionalities, with the exception that in Surveyor (intended for discovery applications), the detection parameters are adjustable, whereas in Explorer (clinical applications), the detection parameters have been set at the default settings in order to achieve analysis to analysis consistency.

## SoftGenetics Technologies and methods used in Mutation Surveyor and Explorer:


**Trace Comparison:** The sample DNA sequence traces are aligned with the reference DNA sequence trace. The DNA nucleotide peak intensities of the traces are normalized to 3000 counts in order to be able to compare the sequence traces of varying intensities. Figure 1 shows the reference sequence trace (top panel), sample sequence traces (middle panel), and mutation electropherogram (third panel) following the normalization and alignment. The analysis results are shown in the mutation electropherogram which has been performed an actual physical trace comparison utilizing our exclusive mathematical anti-correlation algorithm.



**Figure 1.** The top panel shows the DNA sequence traces of the reference. The middle panel shows the sample sequence traces. The third panel is the “difference” between the reference and sample traces. The difference is calculated with our anti-correlation algorithm. A nucleotide mutation is represented as a sharp peak. The four numbers on the mutation peak is the score, mutation peak height, overlapping factor and intensity dropping factor.

### Point Mutation Detection:

The anti-correlation method calculates the difference between the reference and sample traces, noting any found physical differences in the mutation electropherogram. At any specific location, if both the reference and sample are the same type of base, such as the T at the 1550th data point, the anti-correlation will be zero. In the event sequence traces

have different bases at the same point in time, such as the heterozygote change in the sample from G to GA shown in Figure 1, then the anti-correlation will have a very high value. The mutation electropherogram of the anti-correlation profile presents that the two peaks of reference G and sample A are overlapped in space. The mutation electropherogram will present any changes in any of 12 colors corresponding to the possible mutations: AC, AG, AT, CA, CG, CT, GA, GC, GT, TA, TC and TG. To view the color codes click the icon  in the program tool bar.

### **Indel Detection:**

The green center line in the mutation electropherogram represents the monitoring of the sample mobility. The program constantly monitors the sample/patient trace for insertion and deletion differences. In the event a difference is noted the green line will turn red, the software will add a heavy red line over the indel point, and gap either the sample or reference trace until re-alignment is obtained to indicate the inserted or deleted homozygous bases.

### **Advantages of SoftGenetics physical trace comparison technology:**

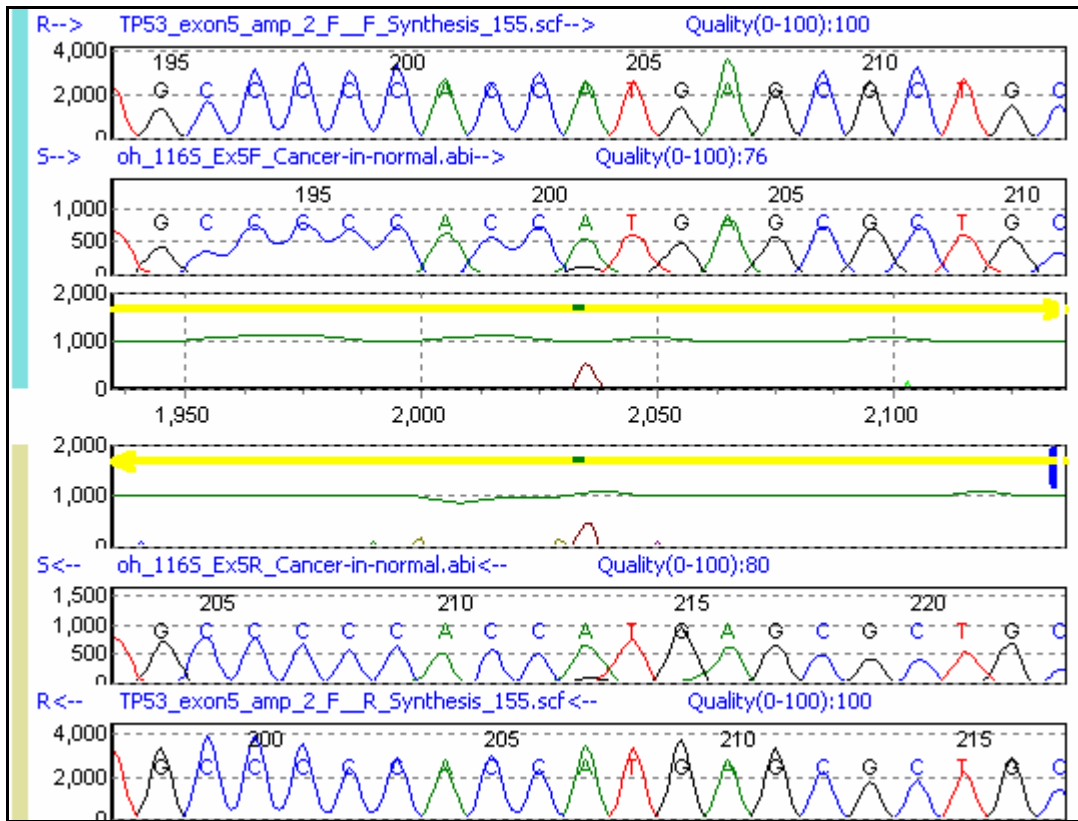
#### **Forgiving to base call errors:**

The anti-correlation technique tolerates alignment errors caused by peak shift. Poor or problematic basecalling errors will not affect the mutation calling. As noted earlier, basecalling accuracy at a Phred score of 18-20 is approximately 98% to 99%; and the rate of human mutations is approximately 1/1200 bases or 0.08%, therefore it is obvious that only an actual trace comparison has the required accuracy. The SoftGenetics algorithm performs an actual physical analysis of the trace, removing any affect of base calling errors.

#### **Increased Sensitivity**

The anti-correlation technique also provides high sensitivity. Figure 2 is an example of sequence sample of the cancer cell population within the large amount of the normal cells. The first three panels represent the reference forward sequence (R→), sample

forward sequence (S→), and forward mutation electropherogram. The second three panels denote the reverse mutation electropherogram, sample reverse sequence (S←) and reference reverse sequence (R←). The mutation peak, G, is under the major peak, A. Our software is able to detect the mutation which is about 10% of populations with the data from both forward and reverse sequences. To activate this level of sensitivity, which may increase false positive rate, click the box: “Check 2Dir Smaller Peaks” by going to software Process: Options : Display : Check 2 Dir Smaller Peaks.



**Figure 2. Sensitivity of the Mutation Surveyor software.** The software detects the smaller portion of the mutation when the forward sequence and reverse sequences are used in the data analysis. The green dots in the centers of mutation electropherogram show the potential mutation. The thick yellow arrow lines indicate the direction of the coding sequence.

## 2. Mutation parameters and mutation scores

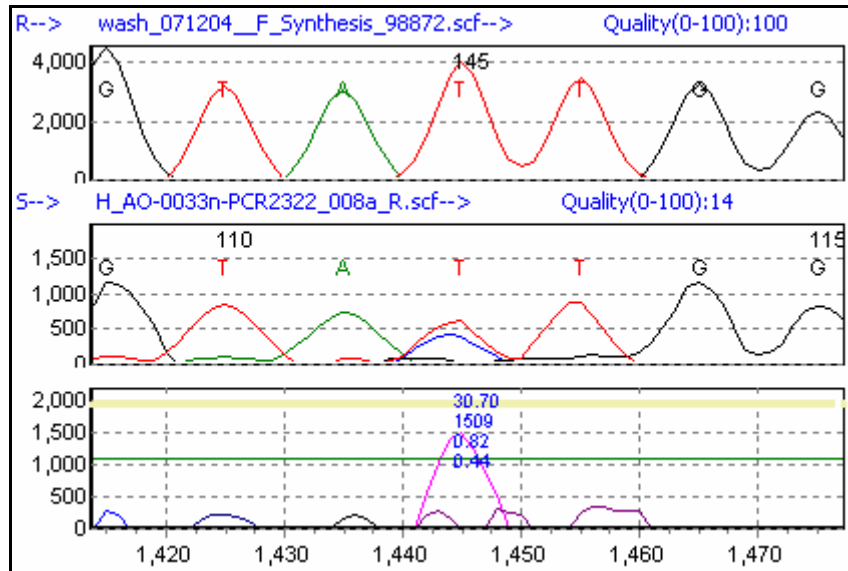
Four parameters are used to discern the mutation peaks that appear in the mutation electropherogram. These four parameters are the mutation height (intensity), overlapping factor, intensity dropping factor, and the signal to noise ratio.

The mutation peak height is the highest peak intensity of the mutation peak in the mutation electropherogram. Mutation noise is the median peak height of all of the mutation smaller peaks in a local section. The first few highest peaks in the mutation electropherogram are rejected. The signal noise ratio is used to determine the confidence of the peaks. The confidence is calculated with Gaussian distribution, assuming that the median value ( $\sigma$ ) is the noise and the highest value is the signal. The area of the Gaussian curve under  $1\sigma$  is 68%,  $2\sigma$  is 95%, and  $3\sigma$  is 99.7%. The error probability of the mutation peak is  $1 - \text{confidence}$ .

$$\text{The mutation score} = -10 \log(\text{error probability}) = -10 \log \left[ \text{erfc} \left( \frac{s/n}{\sqrt{2}} \right) \right]$$

where  $\text{erfc}(x)$  is the complementary error function.

The overlapping factor is calculated with the reference peak to the sample peak of the different color. The overlapping factor is the indicator of relative shift of the two peaks in the horizontal direction.



**Figure 3. Overlapping factor and dropping factor.** The overlapping factor calculates the horizontal (time) overlapping percentage of a reference peak to another sample peak. Within the mutation peak, the reference peak T and sample peak C are overlapped. The dropping factor calculates the relative peak intensity drops at the mutation position.

The dropping factor indicates how much the vertical peak intensity has dropped relative to the neighboring peaks. We have used four peaks to calculate the relative dropping. The two peaks by the mutation are excluded in the calculation, because the mutation

nucleotide often changes the dye binding efficiency of the neighboring peaks. The formula to calculate the dropping factor is

$$= 1 - \frac{4I_{s,mut,x} / I_{r,mut,x}}{I_{s,left1,u} / I_{r,left1,u} + I_{s,left2,v} / I_{r,left2,v} + I_{s,right1,w} / I_{r,right1,w} + I_{s,right2,y} / I_{r,right2,y}}$$

where the three subscripts indicate the sample/reference (s/r), positions such as mutation position (mut) or left1 or right 2, and the type of nucleotides. The above example shows x=T, u=G, v=T, w=G and y=G. The intensity of T in the sample is dropped 44%.

The total mutation score after considering the three parameters using data fusion technology is

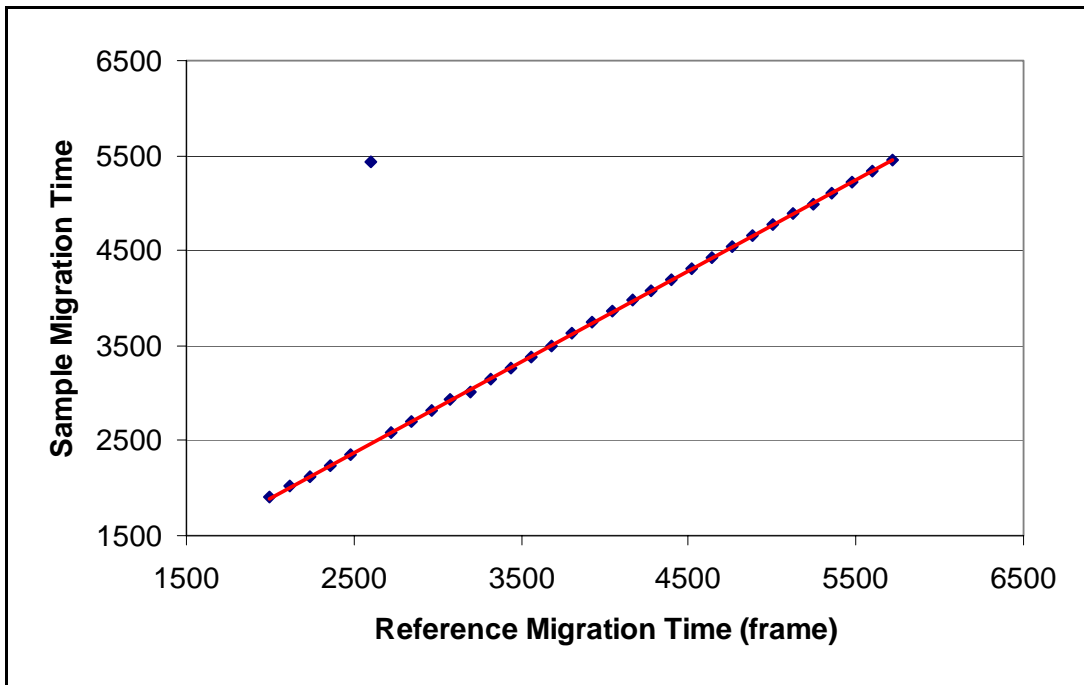
$$= -10 \log \left[ \operatorname{erfc} \left( \frac{s/n}{\sqrt{2}} \right) * \text{dropping\_factor} * \text{overlapping\_factor} \right]$$

When both of the forward and reverse sample sequence traces are compared to the reference, the total mutation score is the summation of the mutation score in both forward and reverse directions.

**Homozygous indel detection and robust alignment.** The software compares the sequence text of the sample sequence to the reference sequence, determining if they are in the same contig. We use the DNA migration time to align the sample trace to reference trace. Three steps are established in the alignment, rough alignment with the sequence text, robust alignment with the DNA migration time, and local adjustment with maximum correlation.

**Rough alignment:** A block of sequence text such as 12 nucleotides (fragment size) in the sample traces is compared to all of the reference trace. If it matches, the program moves to the next 12 nucleotide fragment to find out if next fragment matches the reference. The software requires at least 60 basepairs (the matching base number) to match to the reference. The sample sequences must have at least 30% of the bases matched with the reference sequence, and the 30% is called matching base percentage. These alignment parameters may be adjusted by the user.

**Robust alignment:** When the two sequences are matched with the sequence text. Then the migration time of the reference trace and sample are plotted as x-axis and y-axis, respectively. Then we use the Robust Estimation<sup>15</sup> to find the best linear fit shown in **Figure 4**. The few outliers are ignored by the method. The best fit will be used for the global alignment and then locally. The migration time before and after alignment is captured in the computer memory.



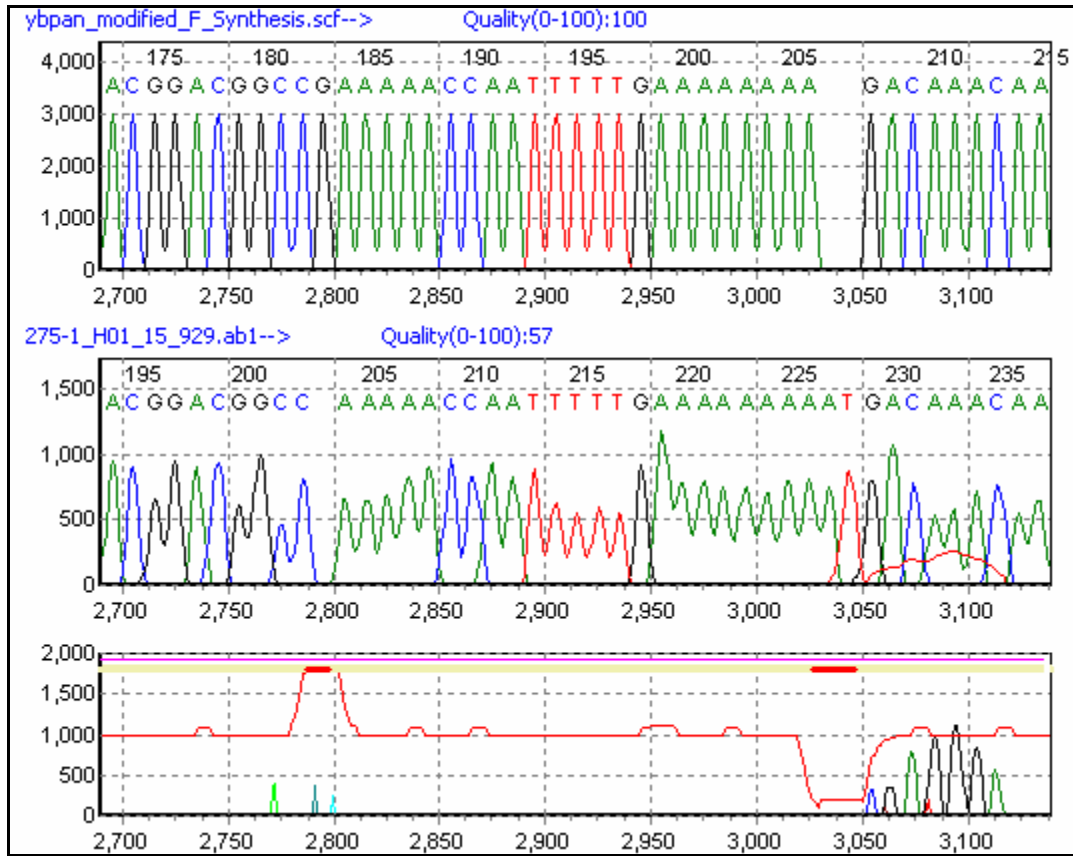
**Figure 4.** The migration time plot of the matching reference and sample traces. A few outliers are ignored by the robust estimation method.

**Local Adjustment:** Maximum correlation and migration time deviation plot. The local sequence alignment is adjusted with the maximum correlation using the mathematical formula:

$$y_{\max} = \sum_i I_{s,i} * I_{r,i}$$

where  $I$  is peak intensity, the subscripts  $s$  and  $r$  standing for the sample and reference. The subscript  $i$  must be the same type of the nucleotide. This formula only calculates the best alignment of the local section in a 12 basepair region to determine the best alignment. It is essential to determine best position of the overall and miscall such as in a reference

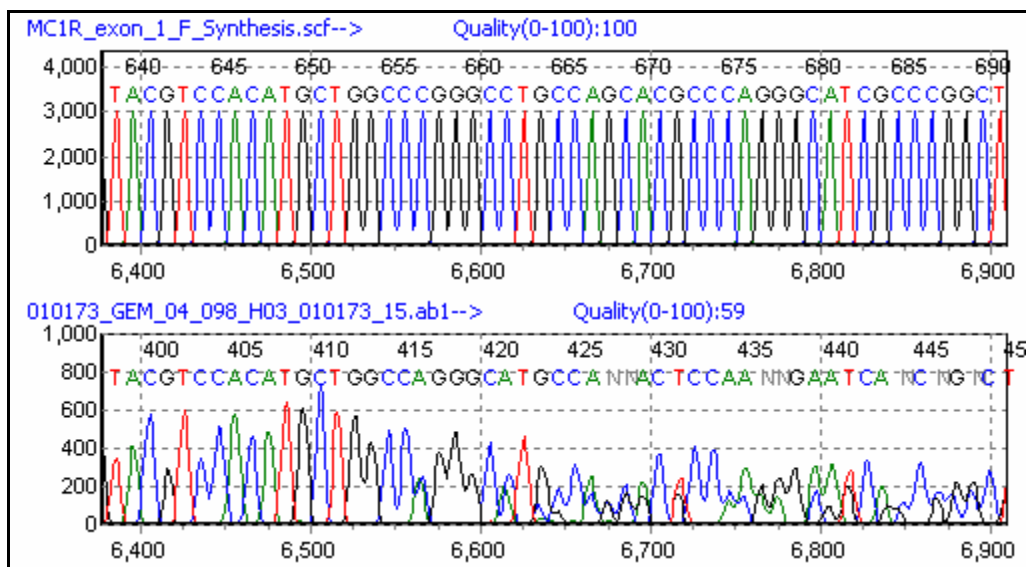
containing “ATTG” and sample ”ATTTG”, the software will then know which T is to align with reference TT. Software will know which T to align with reference TT.



**Figure 5.** Indel detection: the software will “gap either the reference or sample trace to graphically illustrate the indel. The center line will turn from green to red, with heavy red lines above the indel point.

After the sequence is aligned with the above three methods, rough alignment, robust alignment and maximum correlation, a data point of the new time (frame) will correspond a point in the old trace before alignment. The plot of the original 10 data points is altered so that the base pairs before the indel align with the reference file and the base pairs after the indel aligning with the reference files. The resultant plot contains either a gap or an area of compression where the indel occurs. If an insertion occurred, the final plot will contain an area of compression so that the base pairs previous to and following the insertion will still align with the reference file. If a deletion occurs, the final plot will contain a slight gap in the trace because the gap takes place of the deleted base pair so that a constant alignment is maintained.

**4. Heterozygous indel detection.** The DNA sequence traces of the heterozygous mutation sample have a high quality sequence in one portion of the trace and a mixture of two sequence traces after a specific data point. We have developed a method to deconvolute the heterozygous sequence trace into two clean sets of sequence trace with the help of a reference trace or Genbank text file. It is preferable to use a synthetic trace created from the Genbank text file as reference, since the synthetic traces are perfect.

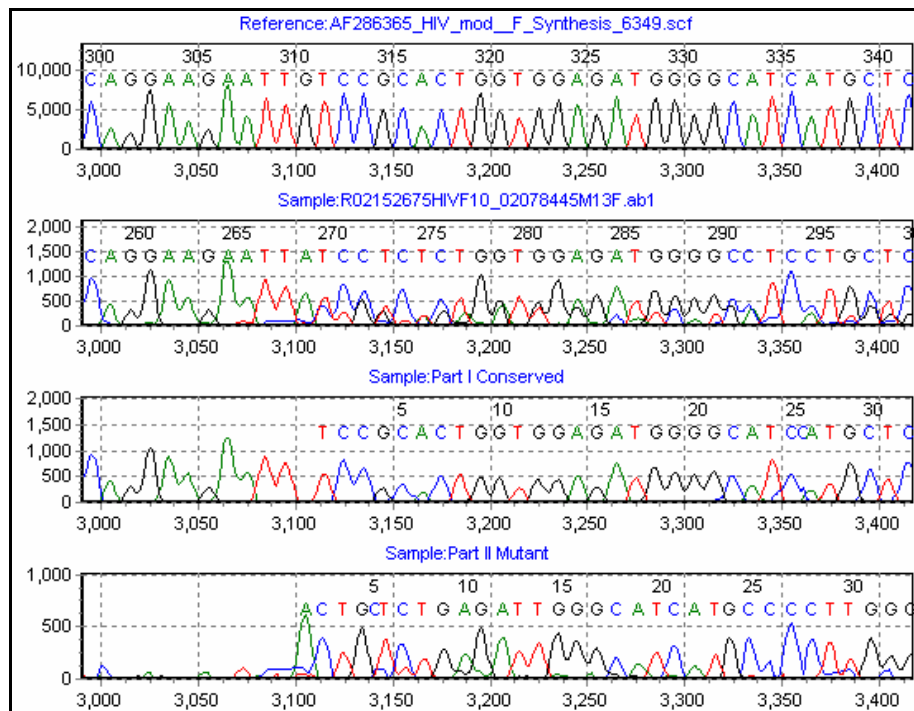


**Figure 6.** The top panel is the reference trace created from the sequence text of MC1R gene found at the NCBI web site. The second panel is the DNA sequence trace from a heterozygous indel sample. The sequence was normal until 415 bp, where patient sequence trace then becomes mixed by the frame shift.

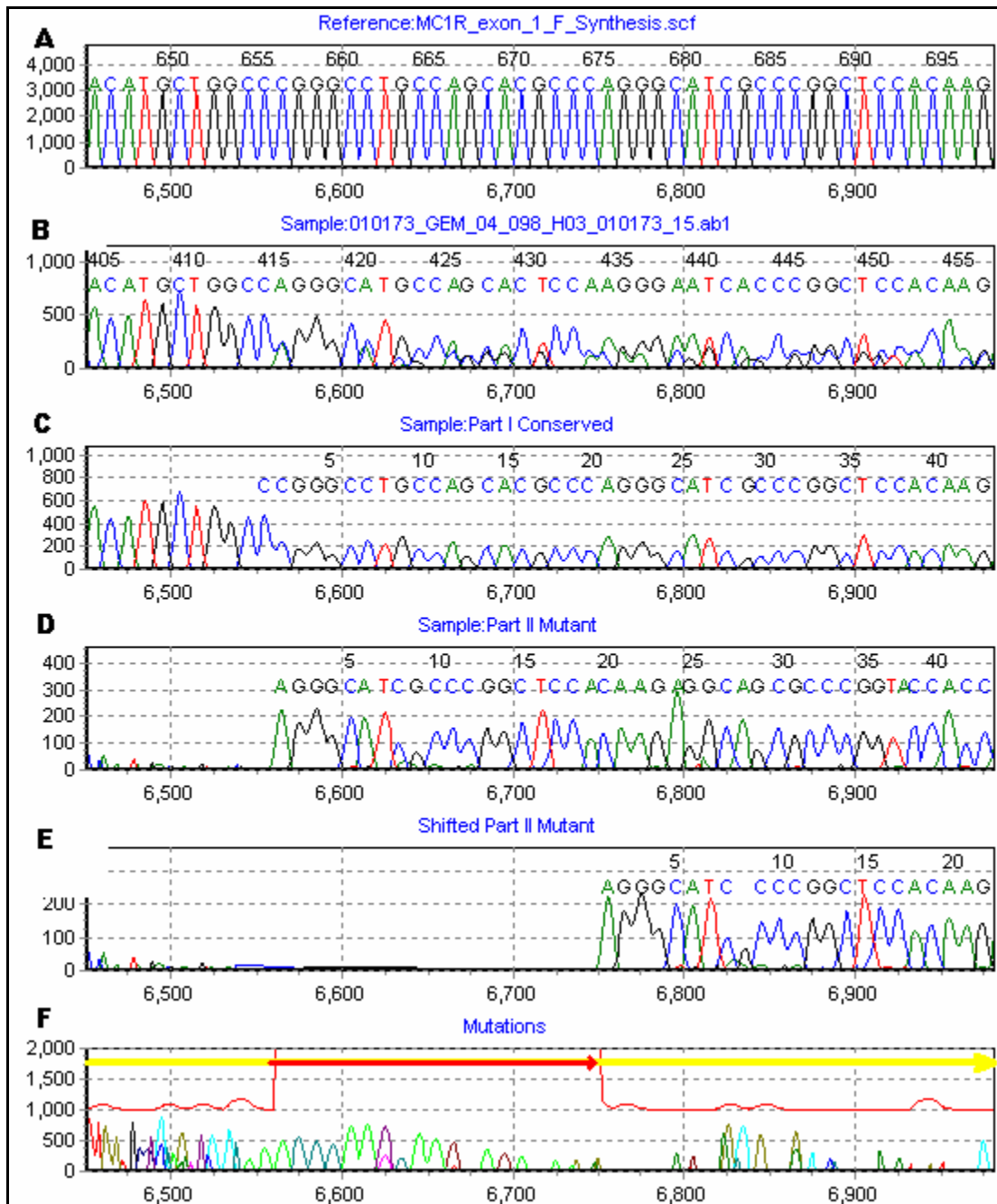
**Sequence Alignment:** The sample sequence trace is aligned with the reference trace by a few iterations. The first step is to align the good quality traces (left portion in Figure 6) using the robust alignment method previously discussed, which gives out the value of the DNA migration time slope. The same slope is extended to the portion of the sequence mixtures (right portion in Figure 6). The sequence trace of the sample is then roughly aligned with the reference traces. Refined alignment is then performed to the traces of the fewer peaks, such as the T bases in Figure 6. The DNA migration time slope adjustment is the key for the iteration process.

**Trace Subtraction:** The reference trace is then subtracted from the sample traces. The intensity ratio is determined from the median peak ratio of the same color at the same location. We prefer to subtract a little more (often 30%) than the determined intensity ratio. This portion is shown in **Figure 8C**. The residual portion (mutant portion) after subtraction is shown in **Figure 7D**. Then we shift the mutant portion of the trace to align with the reference trace again. The heterozygous indel is determined following the migration time adjustment. We have determined in **Figure 7E** that 18 bases have been deleted from 637 to 655 relative to the Genbank, which is marked as 637-655het\_delCGGGCCTGCCAGCACGCC.

**Application:** We are able to deconvolute the trace of **HIV virus** shown in **Figure 9**.



**Figure 7. The heterozygous indel sample of HIV virus.** The HIV virus mutates frequently. Mutation Surveyor software is able to detect the mutations. The data is kindly provided by Filipa Carvalho at Royal Perth Hospital in Australia.



**Figure 8.** Panel A is the reference trace created from the genbank sequence text. B is the samples trace containing the heterozygous indel. C is the conserved trace identical to the reference trace. D is the mutant trace left over after the reference subtraction. E is the shifted trace of the mutant component. We have determined that 18 bases are deleted from 637 to 655 relative to the genbank, which is marked as 637-655het\_delCGGGCCTGCCAGCACGCC.

**Mutation Analysis Reports:** Mutation Surveyor and Explorer both offer a myriad of reporting options. Additionally SoftGenetics will develop one customized report to the user's requirements with program purchase.

Text reports are typically exportable in text (txt), HML or HTML formats.

Sample File	Reference File	Dir	Lane	Gene	Exon	BF	Start	End	Size	Qual	Mut#	Mutation1	Mutation2	Mutation3	Mutation4	Mutation5
1	113a-11.22196.plen	1	1	PTEN	2	11081	11371	291	33	4	11169-77T>C	11169-63T>C	11169-53T>C	11375het_delt	n.a.	
2	113a-11.22196.plen	1	2	PTEN	2	11376	11429	54	0	1	n.a.	n.a.	n.a.	11375het_delt		
3	113a-12.22210.plen	1	1	PTEN	2	11075	11442	368	51	0						
4	113a-12.22210.plen	1	2	PTEN	2	11075	11429	355	58	0						

**Figure 10A. Mutation Report:** Indicates all found variants by lane. Lane cells are linked to analysis electropherogram.

Sample File	Reference File	Dir	Lane	Gene	Exon	BF	Start	End	Size	Qual	Mut#	Mutation1	Mutation2	Mutation3	Mutation4	Mutation5
1	113a-11.22196.plen	1	1	PTEN	2	11081	11371	291	33	4	11169-77T>C	11169-63T>C	11169-53T>C	11375het_delt	n.a.	
2	113a-11.22196.plen	1	2	PTEN	2	11376	11429	54	0	1	n.a.	n.a.	n.a.	11375het_delt		

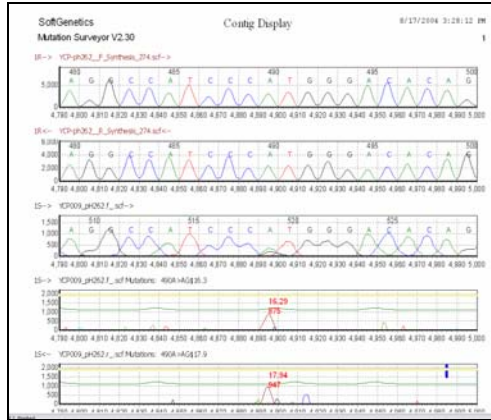
**Figure 10B, Paired Report:** forward & compliments are paired. Variants are listed in variant order across sample set. Allele frequency is supplied at report end.



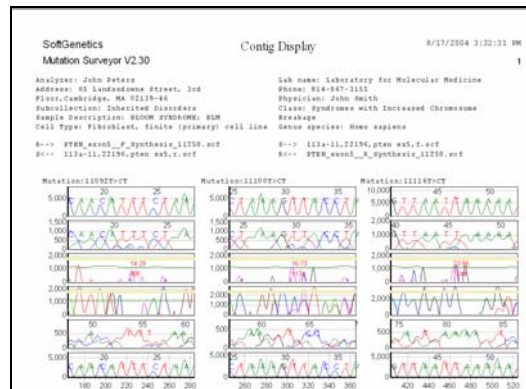
**Figure 11. Paired Global Graphic Display:** Graphically indicates paired compliments, indicating over lap of process able areas, found variant positions, and allele frequency/mutation confidence scoring table.

No	Sample File	Reference File	Dir	Lane	Gene	Exon	BF	Start	End	Size	Qual	Mut#	Mutation1	Mutation2	Mutation3	Mutation4	Mutation5
1	113a-11.22196.plen	113a-11.22196.plen	1	1	PTEN	2	11081	11371	291	33	4	11169-77T>C	11169-63T>C	11169-53T>C	11375het_delt	n.a.	
2	113a-11.22196.plen	PTEN_exon5_R_S1-R	1	1	PTEN	2	11376	11429	54	0	1	n.a.	n.a.	n.a.	11375het_delt		
3	113a-12.22210.plen	PTEN_exon5_F_S1-F	1	1	PTEN	2	11075	11442	368	51	0						
4	113a-12.22210.plen	PTEN_exon5_R_S1-R	1	2	PTEN	2	11075	11429	355	58	0						
5	YCP009_ph262_f	YCP-ph262_F_Synt2-F	1	1	unknown	9	3	13	541	529	90	3	n.a.	245G>AG.G/E/372A>AG.L/L/L400-90A>AG\$1f			
6	YCP009_ph262_r	YCP-ph262_R_Synt2-R	1	2	unknown	9	3	6	499	494	80	4	150-140T>CT\$1	245G>AG.G/E/372A>AG.L/L/L400-90A>AG\$1f			
7	YCP010_ph262_f	YCP-ph262_F_Synt2-F	1	1	unknown	9	3	13	541	529	91	4	n.a.	229C>CT.R/R/(245G>AG.G/E/372A>AG.L/L/L400-90A>AG\$1f			
8	YCP010_ph262_r	YCP-ph262_R_Synt2-R	1	2	unknown	9	3	6	508	503	77	4	229C>CT.R/R/(245G>AG.G/E/372A>AG.L/L/L400-90A>AG\$1f				
9	YCP011_ph262_f	YCP-ph262_F_Synt2-F	1	1	unknown	9	3	13	541	529	89	2	n.a.	372A>G.L/L\$1400-90A>G\$56			
10	YCP011_ph262_r	YCP-ph262_R_Synt2-R	1	2	unknown	9	3	13	499	487	81	2	n.a.	372A>G.L/L\$1400-90A>G\$56			

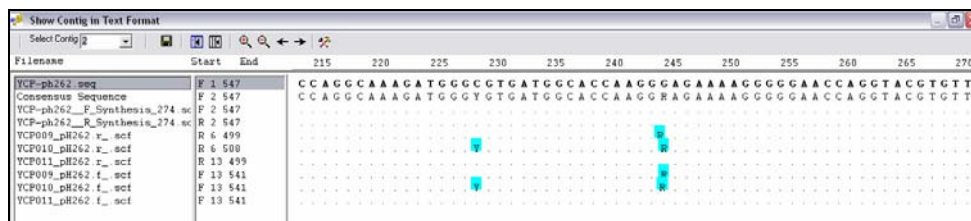
**Figure 12. Advanced two directional report:** Filters out non variant containing lanes



**Figure 13. Printed Analysis Report:** Prints electropherogram of all sample sets, or only mutation containing sample sets.



**Figure 14. Patient Report:** Automatically “cuts & pastes” variant locations with in the patient electropherogram. Header can be customized by individual institution



**Figure 15. Text Report of data set,** available in either consensus or original base call with variants indicated by color code. Previously reported variants are indicated with a colored back ground.

PrettyBase Table						
SNP position	Sample name	Allele 1	Allele 2	IUPAC	Comments	
11092	113a-11.22196.plen.ex5.f.scf	C	T	Y		
11100	113a-11.22196.plen.ex5.f.scf	C	T	Y		
11116	113a-11.22196.plen.ex5.f.scf	C	T	Y		
11375	113a-11.22196.plen.ex5.f.scf		-T			
11373	113a-11.22196.plen.ex5.f.scf		-T			
245	YCP009_pH262_f.scf	A	G	R		
372	YCP009_pH262_f.scf	A	G	R		
490	YCP009_pH262_f.scf	A	G	R		
10	YCP009_pH262_f.scf	C	T	Y		
229	YCP010_pH262_f.scf	C	T	Y		
245	YCP010_pH262_f.scf	A	G	R		
372	YCP010_pH262_f.scf	A	G	R		
490	YCP010_pH262_f.scf	A	G	R		
503	YCP010_pH262_f.scf	A	A	A	?	
504	YCP010_pH262_f.scf	G	T	K	?	
372	YCP011_pH262_f.scf	G	G	G		
490	YCP011_pH262_f.scf	G	G	G		

**Figure 16. Directly Exportable “pretty base” format report, reports found variants by base position and in IUPAC terminology.**

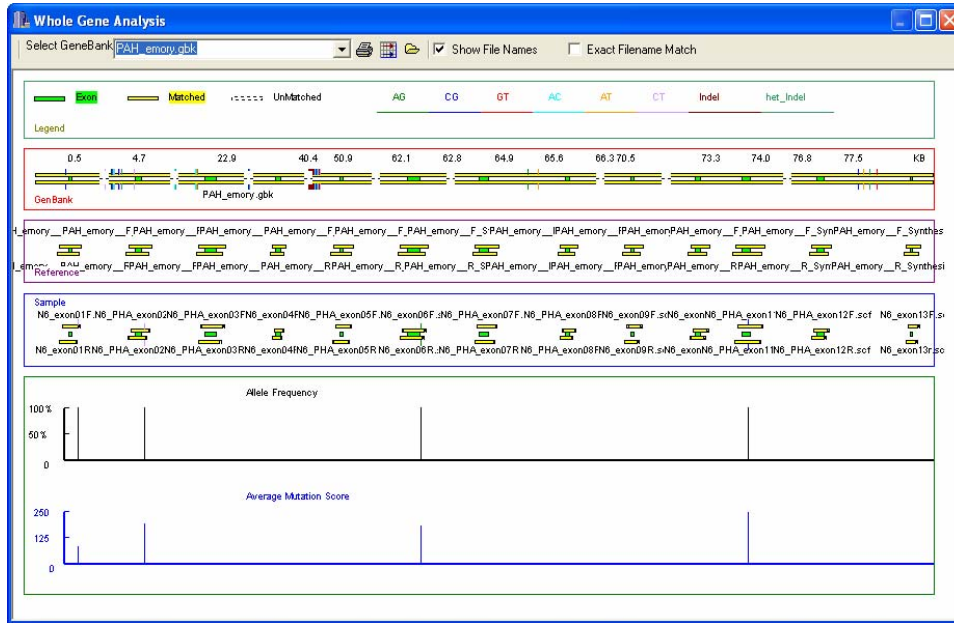
**Whole gene analysis and mutation assembly:** Mutation Surveyor software will assemble multiple mutation projects into a whole gene analysis project. The users may input the n\_primer match text files or 2D\_match text file.

Mutation Surveyor has incorporated yet another very useful function into the already outstanding software package. The “Open Whole Gene Data...” option allows users the ability to fuse multiple SoftGenetics Project (.sgp) files into a single sample analysis. This function permits the use of multiple primer sets to be incorporated into analysis of a single larger sample.

The user has the option of loading Genbank or Sequence files as a reference for their analysis. Then the user can select as many SoftGenetics Project Files (.sgp) as they desire to be loaded into the sample analysis. The use of a text file is an allowed option for sample matching. A spreadsheet program such as Excel should be used, with each of the desired file names for the analysis grouped horizontally in the same row. The filenames for each analysis can be added to a matrix and grouped by rows, with each row containing the file names to be included in the specific analysis.

In this specific type of analysis, certain portions of sequences with different primers may possibly overlap each other. The regions in which the sequences overlap serve as internal controls. When such an overlap occurs, the mutation detection in that specific region becomes extremely accurate. This is due to the fact that if a mutation truly exists, it should be present in a specific region, regardless of the primer set. For example, if a mutation is found in at a distinct location with one specific primer set, and

then a mutation was found at the same point with another primer which overlapped the first, we can be very confident that the mutation actually exists because of the superior accuracy of these overlapping regions. Furthermore, if a mutation is found at a distinct location within one specific primer set, and then a mutation was not set with another primer which overlaps the first, we can be fairly confident that no mutation exists at this point.



**Figure 17. Whole Gene Analysis** The software will align multiple fragments to the reference, and then calculate an allele frequency with mutation confidence score at each found variant position.

## Reference and Software Citation Papers

- <sup>1</sup> Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, Markowitz S, Willson JK, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. 2003. Mutational analysis of the tyrosine kinome in colorectal cancers. **Science**. 300(5621):949.
- <sup>2</sup> Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell SM, Riggins GJ, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE. 2004. High frequency of mutations of the PIK3CA gene in human cancers. **Science**. 304(5670):554.
- <sup>3</sup> Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. 2004. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. **Science**. 304(5676):1497-500.
- <sup>4</sup> Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS, Parmigiani G, Yan H, Wang TL, Riggins G, Powell SM, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE. 2004. Mutational analysis of the tyrosine phosphatome in colorectal cancers. **Science**. 2004 304(5674):
- <sup>5</sup> Nickerson DA, Tobe VO, Taylor SL, 1997  
PolyPhred: automating the detection and genotyping of single nucleotide substitutes by fluorescence-based sequence of PCR products.  
**Nucleic Acid Research**, 25, 2745-2751.
- <sup>6</sup> Bonfield JK, Rada C, Staden R, 1998.  
Automatic detection of point mutation using fluorescent sequence trace subtraction.  
**Nucleic Acid Research**, 26, 3404-3409.
- <sup>7</sup> Rajagopalan H, Jallepalli PV, Rago C, Velculescu VE, Kinzler KW, Vogelstein B, Lengauer C. 2004. Inactivation of hCDC4 can cause chromosomal instability. **Nature**. 428(6978):77-81.
- <sup>8</sup> Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE. 2002 Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. **Nature**. 29;418(6901):934.
- <sup>9</sup> Rosenberg EH, Almeida LS, Kleefstra T, deGrauw RS, Yntema HG, Bahi N, Moraine C, Ropers, RH, Fryns JP, deGrauw TJ, Jakobs C, Salomons GS, 2004,  
High Prevalence of SLC6A8 Deficiency in X-Linked Mental Retardation Identifiers  
*The American Journal of Human Genetics*, 75 (97).
- <sup>10</sup> Fakhrai-Rad H, Zheng J, Willis TD, Wong K, Suyenaga K, Moorhead M, Eberle J, Thorstenson YR, Jones T, Davis RW, Namsaraev E, Faham M, 2004  
SNP Discovery in Pooled Samples With MismatchRepair Detection.  
*Genome Research* 1404-1412
- <sup>11</sup> Shulenin S, Nogee LM, Annilo T, Wert SE, Whitsett JA, Dean M. 2004,  
ABCA3 gene mutations in newborns with fatal surfactant deficiency.  
*N Engl J Med*. 2004 Mar 25;350(13):1296-303.
- <sup>12</sup> Landi MT, Goldstein AM, Tsang S, Munroe D, Modi W, Ter-Minassian M, Steighner R, Dean M, Metheny N, Staats B, Agatep R, Hogg D, Dalista DC, 2004  
Genetic susceptibility in familial melanoma from northeastern Italy.  
*J. Med Genetics*, 2004;41:557-566.
- <sup>13</sup> [http://www.appliedbiosystems.com/support/download/seqscape//readme/intro\\_to\\_seqscapev2.1.1\\_demo.pdf](http://www.appliedbiosystems.com/support/download/seqscape//readme/intro_to_seqscapev2.1.1_demo.pdf)
- <sup>14</sup> <http://www.paracel.com/sas/agent.htm>
- <sup>15</sup> Press W.R. et al, Numerical recipe in C, Chapter 16, Modeling of data, Page 699.