Analysis of sequence data in the cancer genome project

Patrick Tarpey

Team 78

Welcome Trust Sanger Institute

Wellcome Trust Sanger Institute

Use the human genome sequence and high throughput mutation detection techniques to identify somatically acquired sequence variants/mutations and hence identify genes critical in the development of human cancers

Wellcome Trust
Sanger
Institute

Use the human genome sequence and high throughput mutation detection techniques to identify somatically acquired sequence variants/mutations and hence identify genes critical in the development of human cancers

The detection of large homozygous deletions provides information that may lead to the isolation of recessive oncogenes which have been inactivated by the deletion.

Small intragenic mutations are commonly found in both recessive oncogenes and dominantly acting oncogenes This project is systematically screening coding exons and flanking splice junctions of all genes in the human genome for somatically acquired small intragenic mutations in human cancer

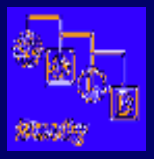
Wellcome Trust Sanger Institute

Use the human genome sequence and high throughput mutation detection techniques to identify somatically acquired sequence variants/mutations and hence identify genes critical in the development of human cancers

The detection of large homozygous deletions provides information that may lead to the isolation of recessive oncogenes which have been inactivated by the deletion.

Small intragenic mutations are commonly found in both recessive oncogenes and dominantly acting oncogenes This project is systematically screening coding exons and flanking splice junctions of all genes in the human genome for somatically acquired small intragenic mutations in human cancer

X Project



Genetics of learning Disability Aims to identify novel genes involved in non-syndromic X-linked mental retardation

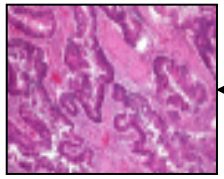
Wellcome Trust Sanger Institute

CGP

518 Genes
10,000 STS's

Set of Tumour samples and normals

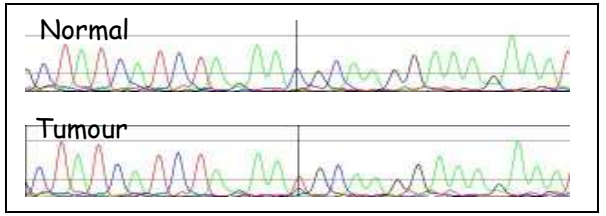
Colorectal



PCR genes

Resequence

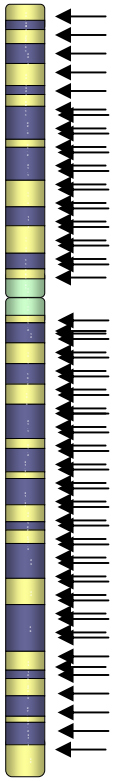
Analysis



Somatic mutations

Confirmation work

X project-candidate screen



Candidate screen

100 candidate genes

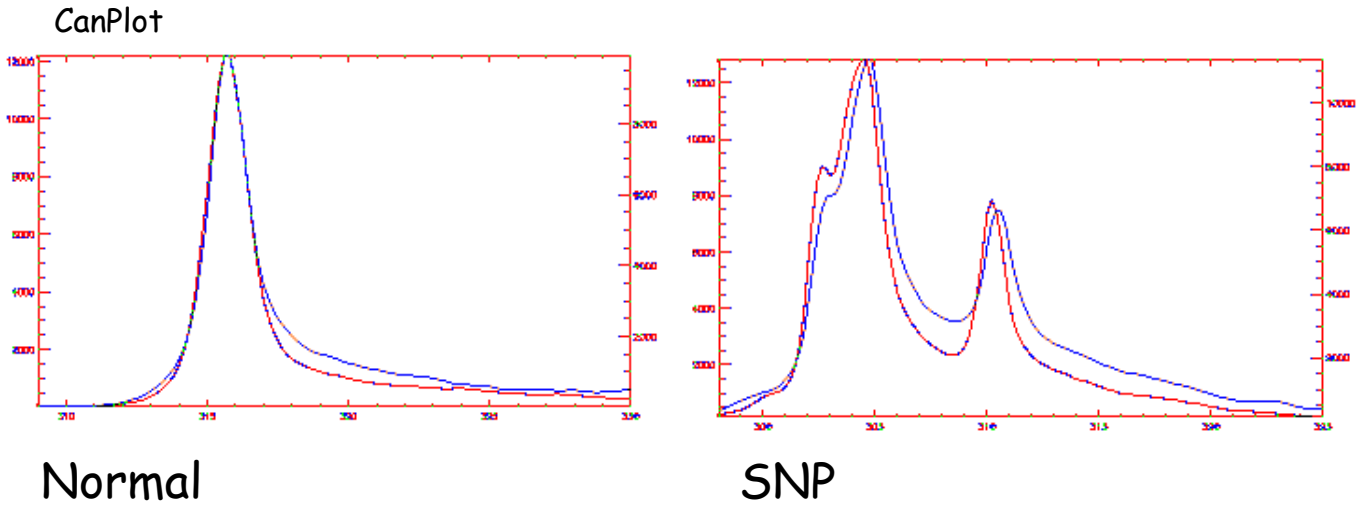
~400 MR families

- High-throughput mutation detection
  - ~80-100 000 PCR's week
- Somatic heterozygous mutations in tumours can be subtle
  - Insertions/deletions
  - Aneuploidies
  - Normal contamination

Mutation detection of small intragenic variants

Conformation sensitive capillary electrophoresis (CSCE)

Cancer
Blood



Normal

SNP

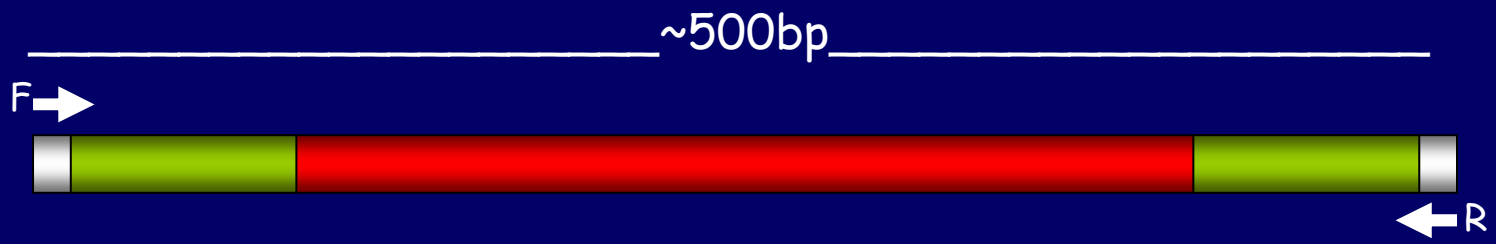
Mutation detection of small intragenic variants

- **Re-sequencing**
  - More robust at high throughput (fewer fails)
  - Affordable
  - Quick
  - Automated data analysis

  - 10 ABI 3730's

Wellcome Trust Sanger Institute

PCR protocol



- Amplimers are a maximum of 500 bp (3730 rapid runs, 30 min)

Wellcome Trust Sanger Institute

PCR protocol

- Amplimers have ~75bp 'buffer Zone' including primer

PCR protocol

____Region of interest Exon +/- 6bp____

Wellcome Trust Sanger Institute

PCR protocol

____Region of interest Exon +/- 6bp____

PCR protocol

Two temperatures

Single buffer

Purified

ExoSAP

Sequencing protocol

BigDye (1/16 diln)

Purification
Ethanol ppt

Resuspension
Water

Wellcome Trust Sanger Institute

Data analysis

BLAT

In house (homozygous variants)

Mutation Surveyor

CSA

In house development

Wellcome Trust Sanger Institute

BLAT

- Phred basecalls AB1/SCF files
-
- Aligns to genome

- Identifies differences in basecall alignment
  - (assess quality using phred quality scores)

- Restricts analysis to ROI and excludes known SNP's

- Compares identified variants in forward and reverse strands

- Records coverage

Wellcome Trust Sanger Institute

BLAT

Text output

312	3972	r	g,98,a,46129870	2_strands	OPPOSITE HAS VARIANT
312	3972	f	g,170,a,46129870	2_strands	OPPOSITE HAS VARIANT
370	3977	f	t,59,c,46131655	2_strands	OPPOSITE HAS VARIANT
370	3977	r	t,190,c,46131655	2_strands	OPPOSITE HAS VARIANT
016	3972	f	1,94,-,46129796	1_strand	
314	3974	r	c,169,a,46130381	1_strand	
050	3974	r	c,166,a,46130381	1_strand	

BLAT

Advantages

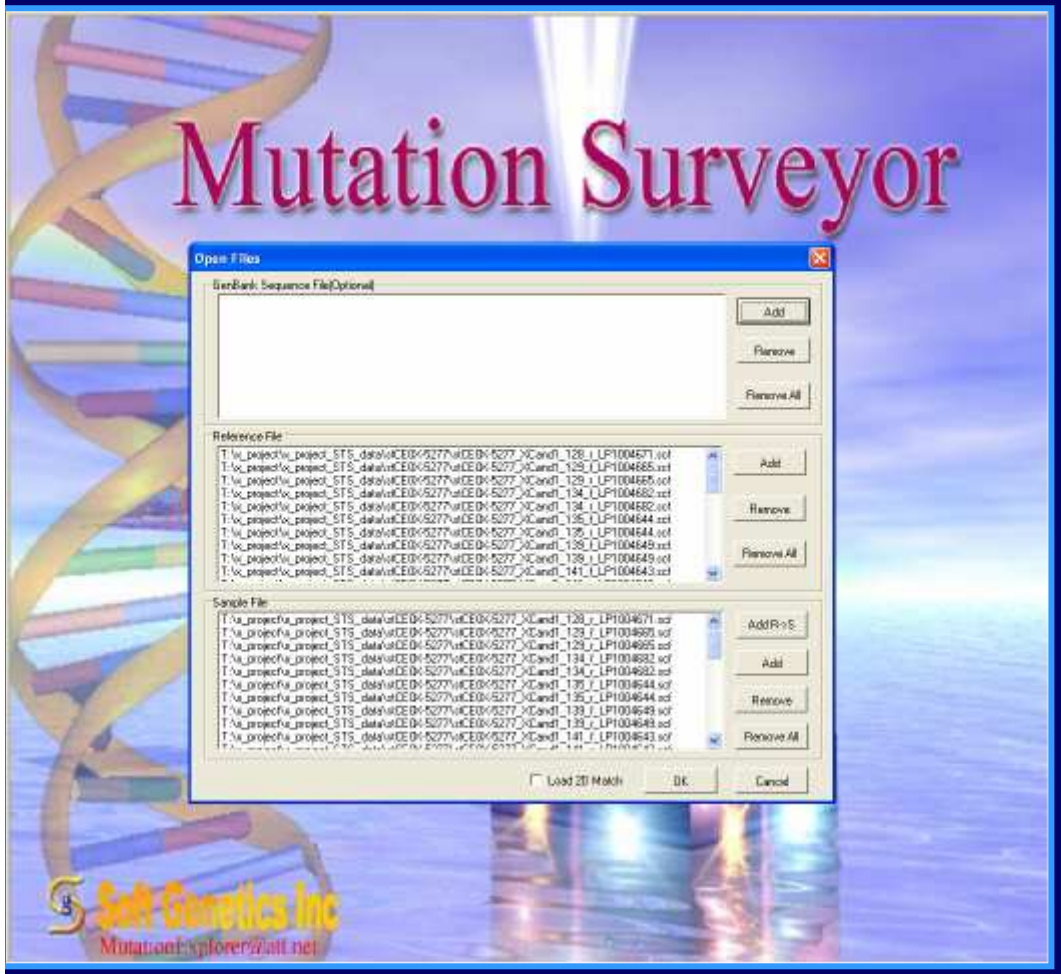
- Simultaneous analysis of multiple STS's in multiple samples
- Rapid analysis
- Few false positives
- Inputs sequence files only

BLAT

Disadvantages

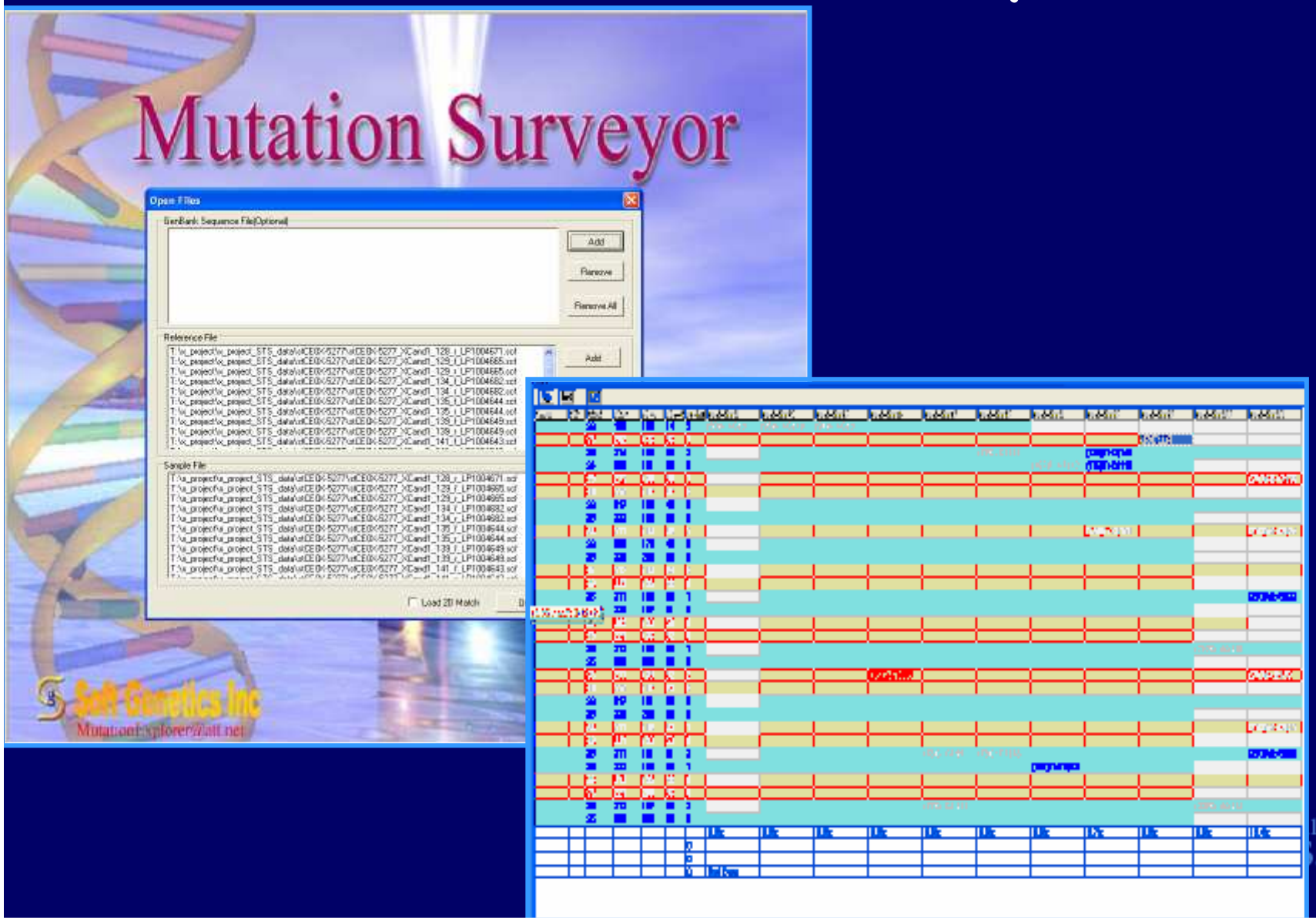
- Only detects homozygous variants
- Trace verification of variants required

Mutation Surveyor

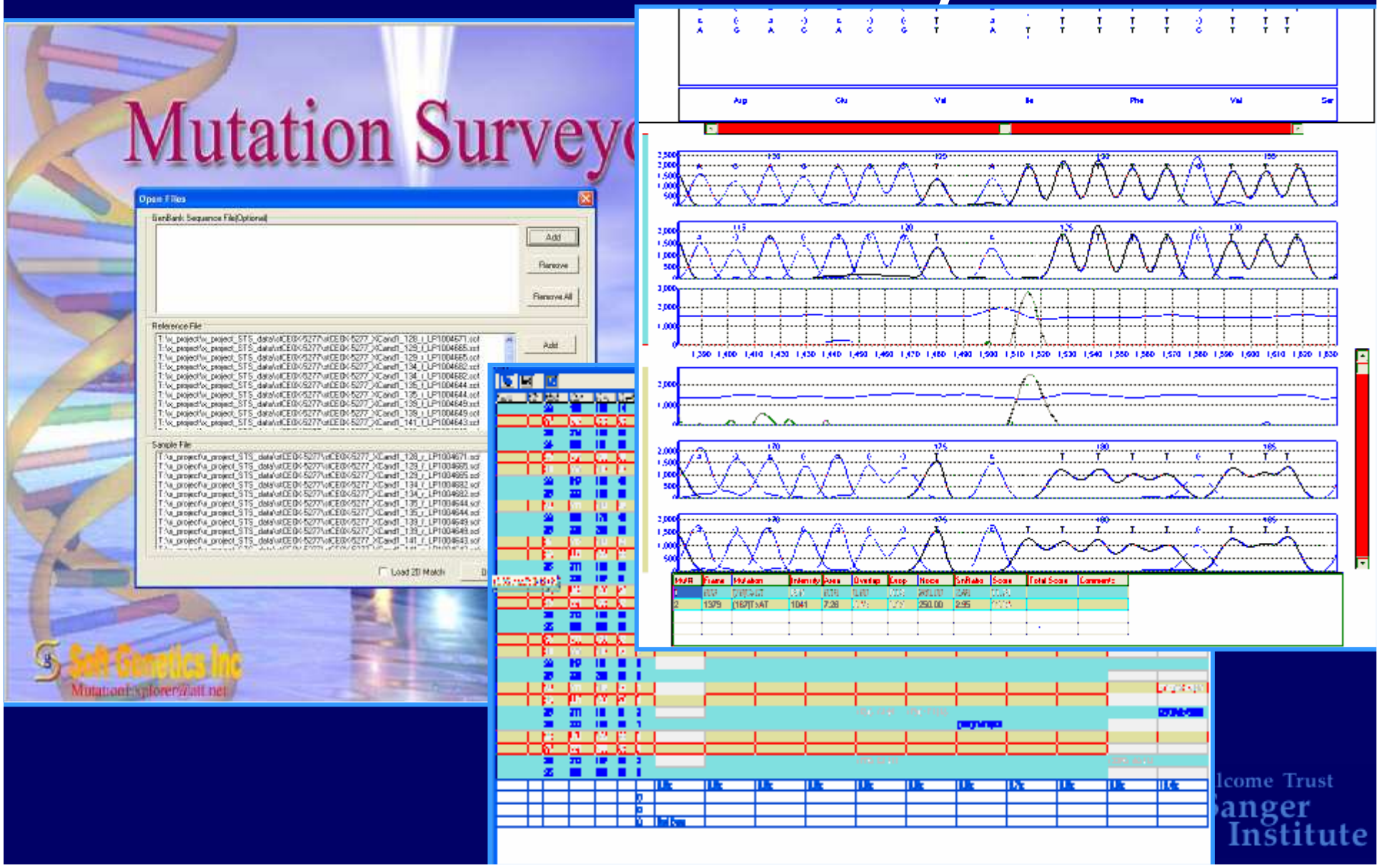


Wellcome Trust Sanger Institute

Mutation Surveyor



Mutation Surveyor



lcome Trust
anger
Institute

Mutation Surveyor

- 5 Mutation parameters
  - Mutation height
    - Must exceed minimum value
  - Overlapping factor
    - Detects presence of mutant peak
  - Dropping factor
    - Detects drop in height of wild type peak compared to reference peak (normalised with 2 upstream and 2 downstream peaks)
  - SN-ratio
    - Signal: mutation peak intensity
    - Noise: Median peak intensity around mutant peak
  - Mutation score
    - Mutation error probability score
    - Noise/overlapping factor/dropping factor

Wellcome Trust Sanger Institute

Mutation Surveyor

- Advantages

  Quick

  High-throughput

  Simple interfaces

  Identifies variants in reference Electropherogram

  ROI

  Coverage

Wellcome Trust Sanger Institute

Mutation Surveyor

- Disadvantages

  Sensitivity?

  homozygous variants
  heterozygous insdels

  Intensity information lost

  (? Overloaded)

  Inflexible

Wellcome Trust Sanger Institute

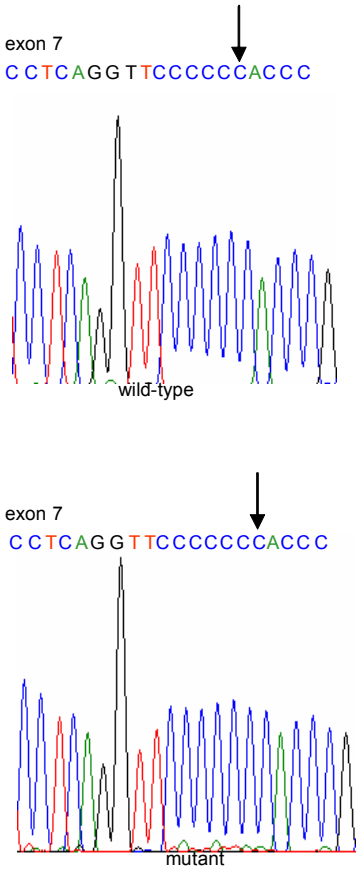
Mutation Surveyor

- Disadvantages

  Sensitivity?

    homozygous variants
    heterozygous insdels

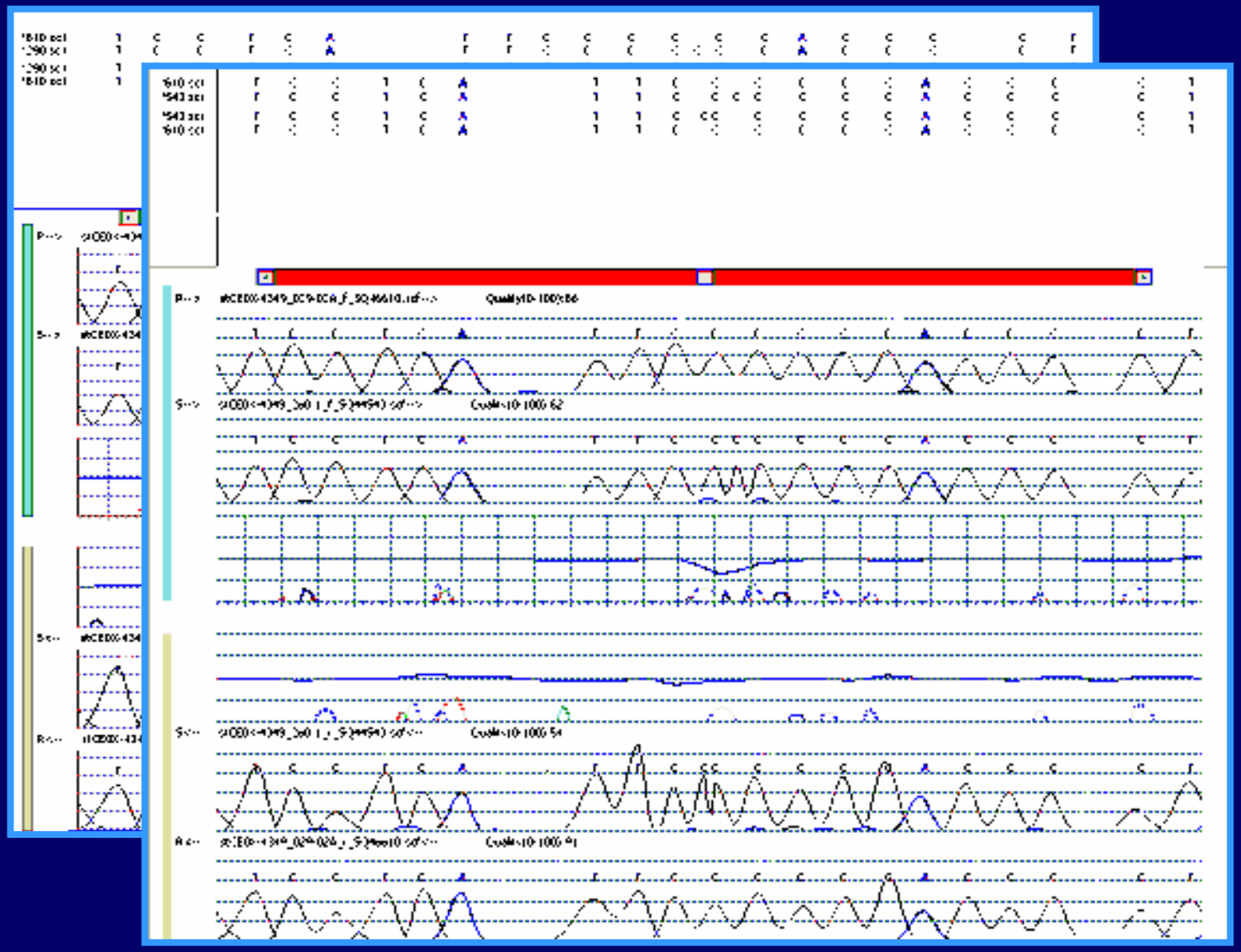
DLG3, 1087insC



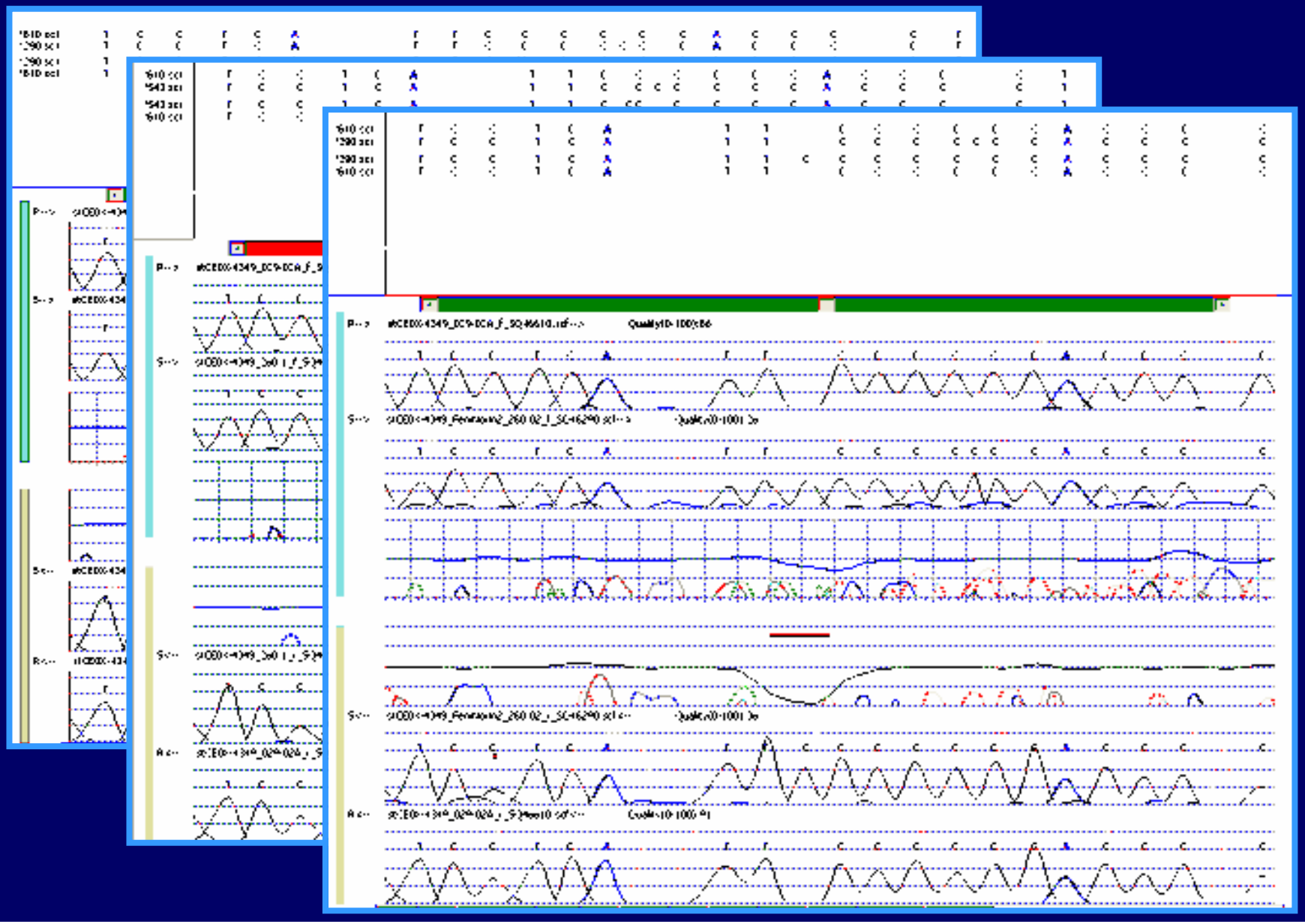

```
841 GTGAACAACACCAATCTGCAGGATGTGAGGCACGAGGAAGCTGTGGCCTCACTGAAGAAC  PDZ2
281 -V--N--N--T--N--L--Q--D--V--R--H--E--E--A--V--A--S--L--K--N-

901 ACATCTGATATGGTGTATTTGAAGGTGGCCAAGCCAGGCAGCCTCCACCTCAACGACATG
301 -T--S--D--M--V--Y--L--K--V--A--K--P--G--S--L--H--L--N--D--M-

961 TACGCTCCCCCTGACTACGCCAGCAC|TTTTACTGCCTTGGCTGACAACCACATAAGCCAT
321 -Y--A--P--P--D--Y--A--S--T--F--T--A--L--A--D--N--H--I--S--H-
                              7

1021 AATTCCAGCCTGGGTTATCTCGGGGCTGTGGAGAGCAAGGTCAGCTACCCTGCTCCTCCT
 341 -N--S--S--L--G--Y--L--G--A--V--E--S--K--V--S--Y--P--A--P--P-

1081 CAGGTTCCCCCCCACCCGCTACTCTCCTATTCCCAGGCACATGCTGGCTGAGGAGGACTTC
 361 -Q--V--P--P--H--P--L--L--S--Y--S--Q--A--H--A--G--X--E--D--F-
        |8
1141 ACCAGAGAGCCTCGCAAGATCATCCTGCACAAAGGCTCCACAGGCCTGGGCTTCAACATC
 381 -T--R--E--P--R--K--I--I--L--H--K--G--S--T--G--L--G--F--N--I-
```

Mutation Surveyor

Mutation Surveyor



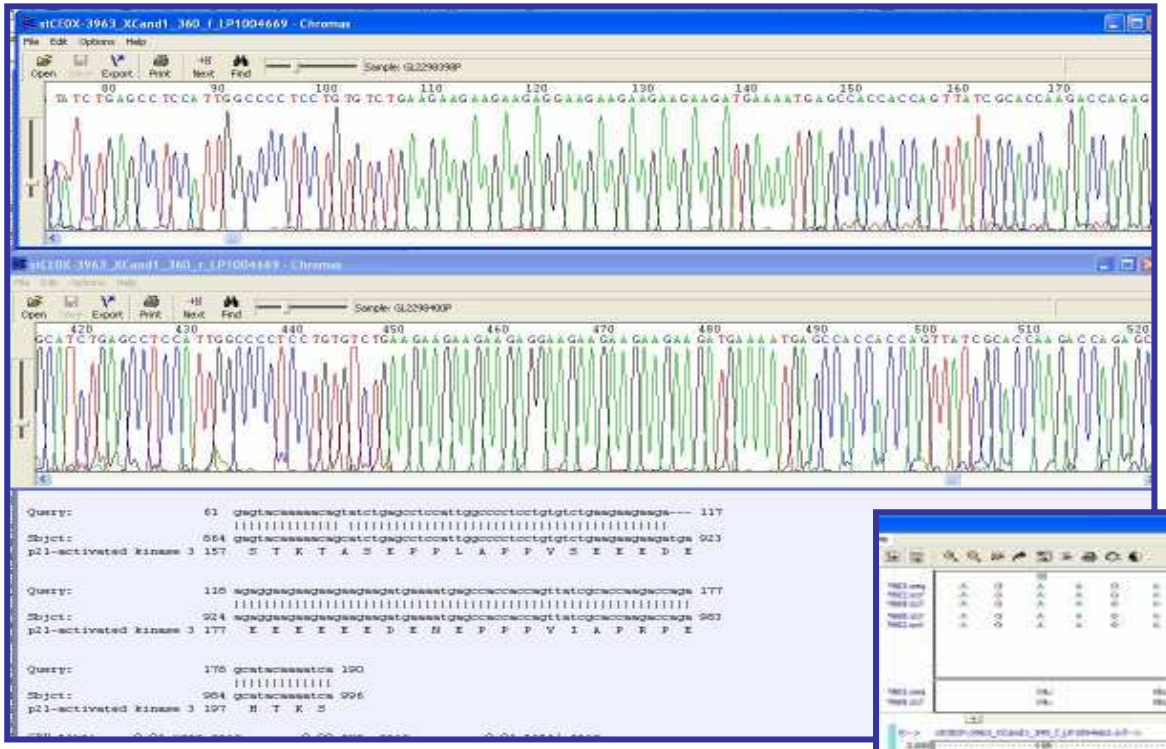
Mutation Surveyor

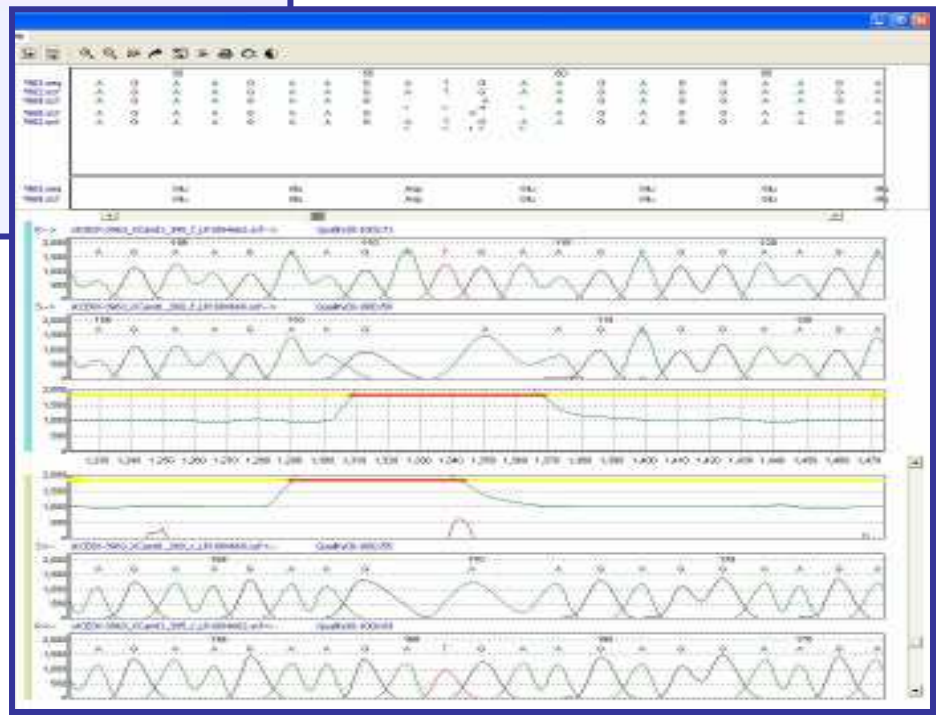


Homozygous mutations

Sample No.	STS	Mutation	Mutation		BLAT	Surveyor
			Nucleotide	amino acid		
312	3972	Missense	A703>G	N103>D	✔	✔
370	3977	Splice	-5C>T		✔	✔
324	5103	Missense	C1953>T	T625>M	✔	✔
385	3960	Frameshift	526delA		✔	✔
134	5277	Frameshift	839insA		✔	X
360	3963	in frame deletion	921-923delTGA	175del D	✔	X
393	3964	Missense	A1103>G	N236>S	✔	✔
135	5258	Missense	G1220>C	G275>A	✔	✔
72	4077	in frame deletion	454-456delGGA	152delG	✔	X
327	4074	Silent	A214>G	L21>L	✔	✔
123	3600	Missense	T2767>C	W914>R	✔	✔
16	3584	Missense	T829>C	Y268>H	✔	✔
74	3596	Missense	A2156>G	D710>G	✔	✔

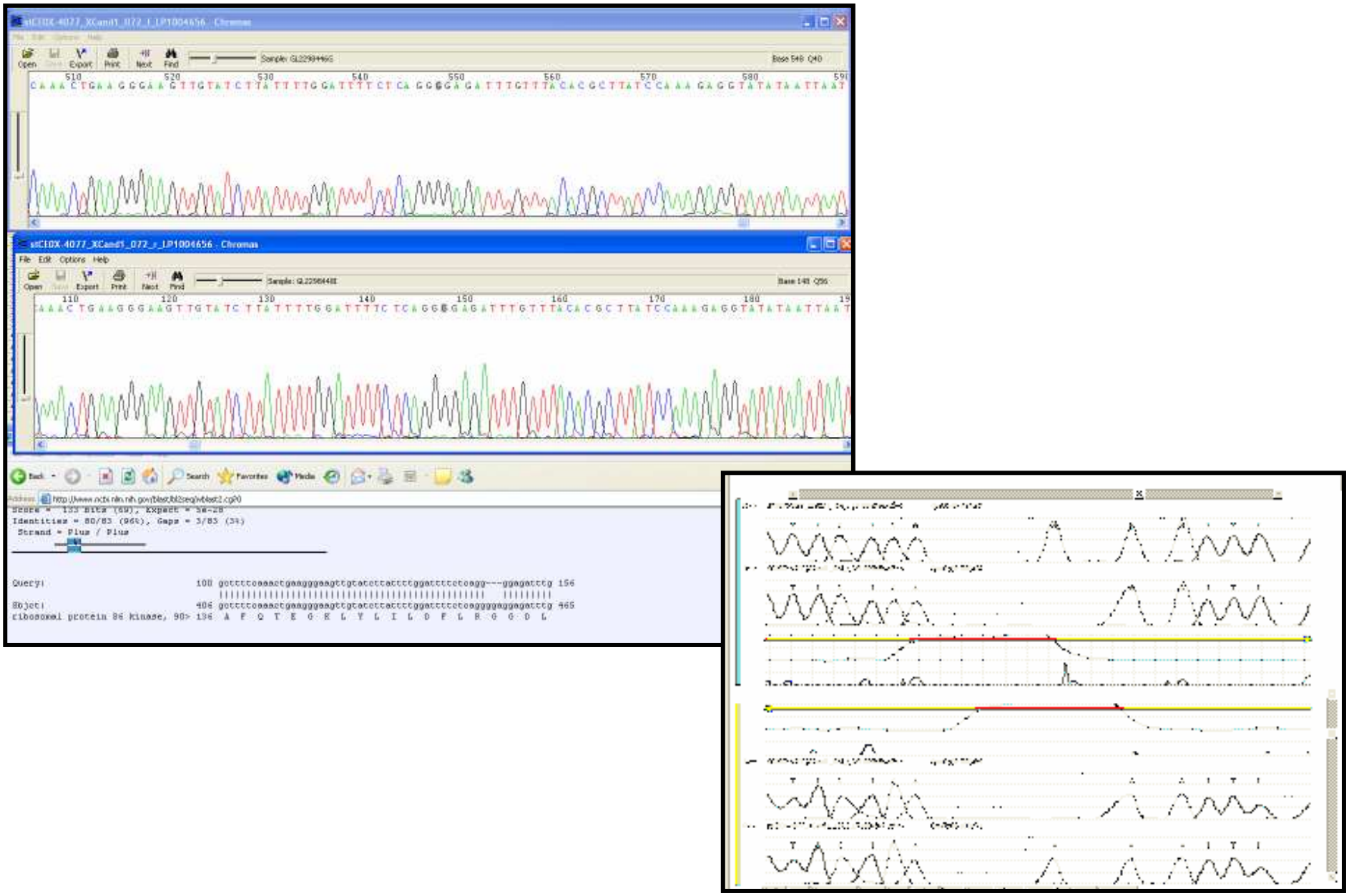
Del 3bp





insA

Del 3bp



Development-CSA

Comparative sequence analysis

Comparative analysis of normalised peak heights in a sample trace compared to a reference trace

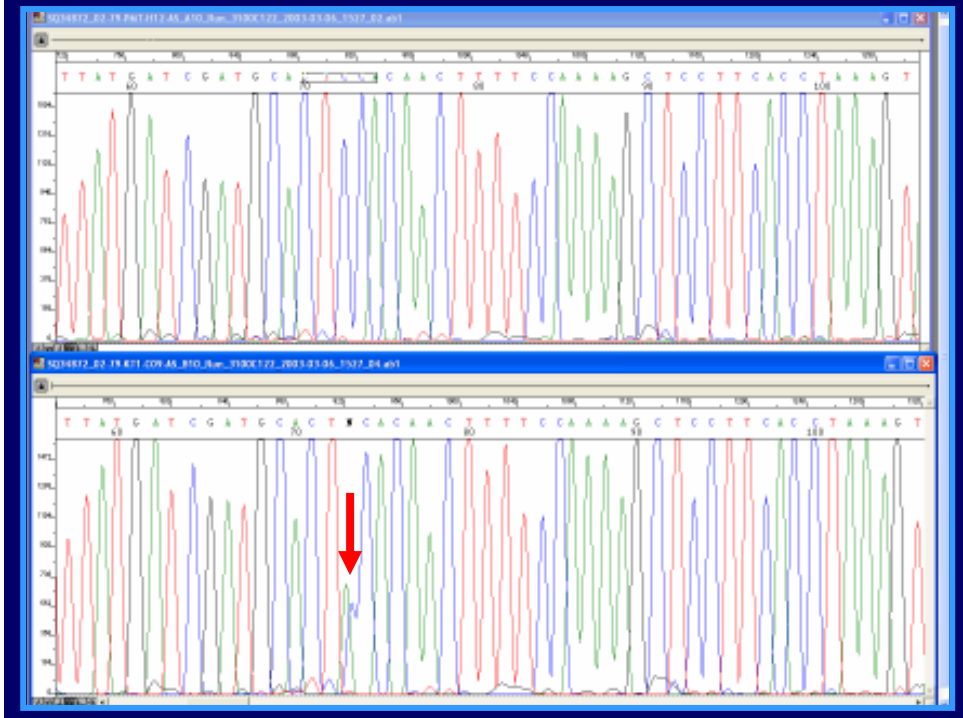
Analysis performed on raw data

Auto-CSA (Excel)

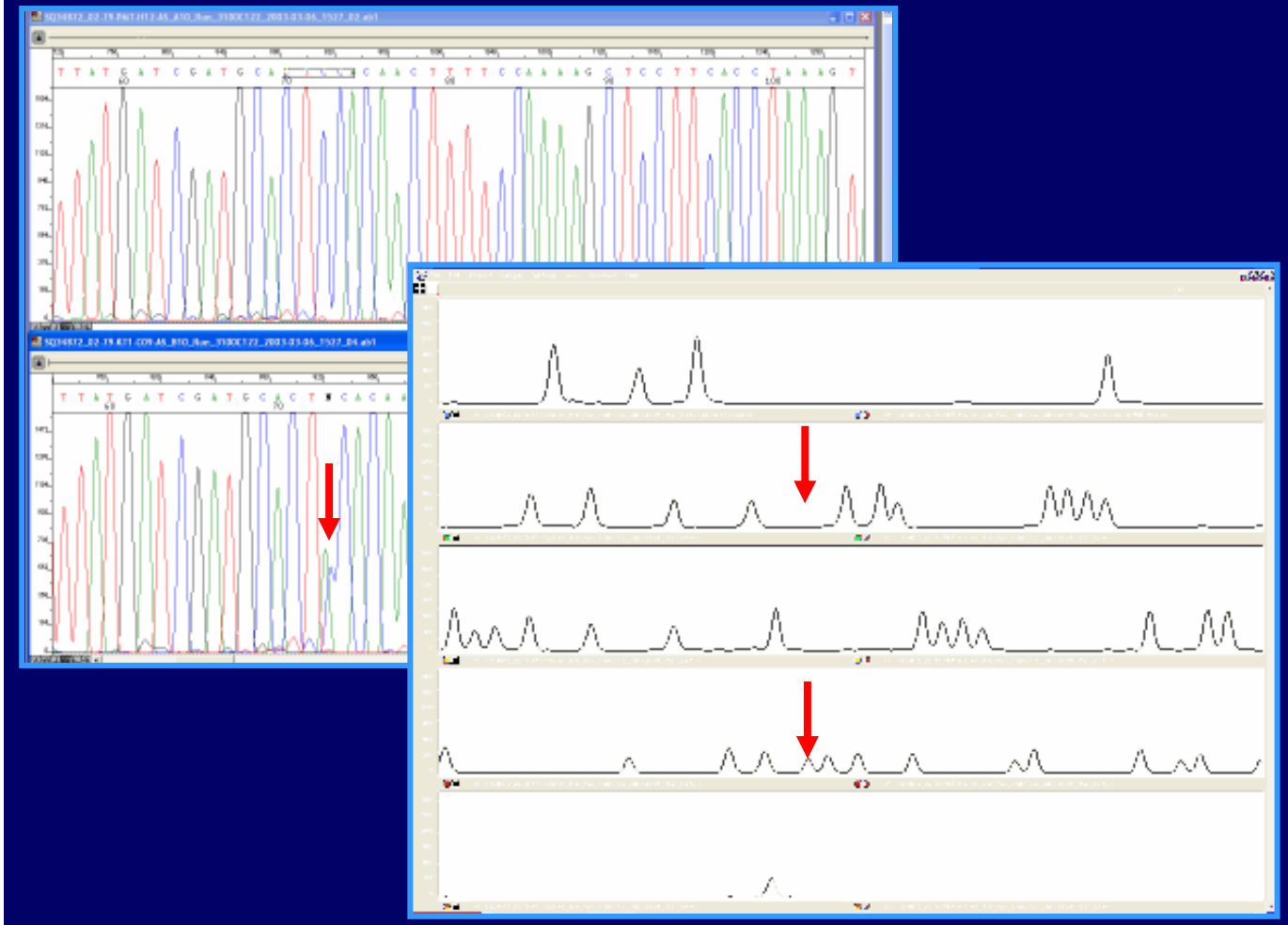
Automated-comparative sequence analysis

- Use Genescan/Genemapper to identify and quantify peaks.
- Use autoCSA to extract information for the real peaks and ignore noise peaks (using reference text)
- Compare normalised peak heights between sample and reference.
- If a normalised peak drops by >35%, ? Mutation (trace verify)

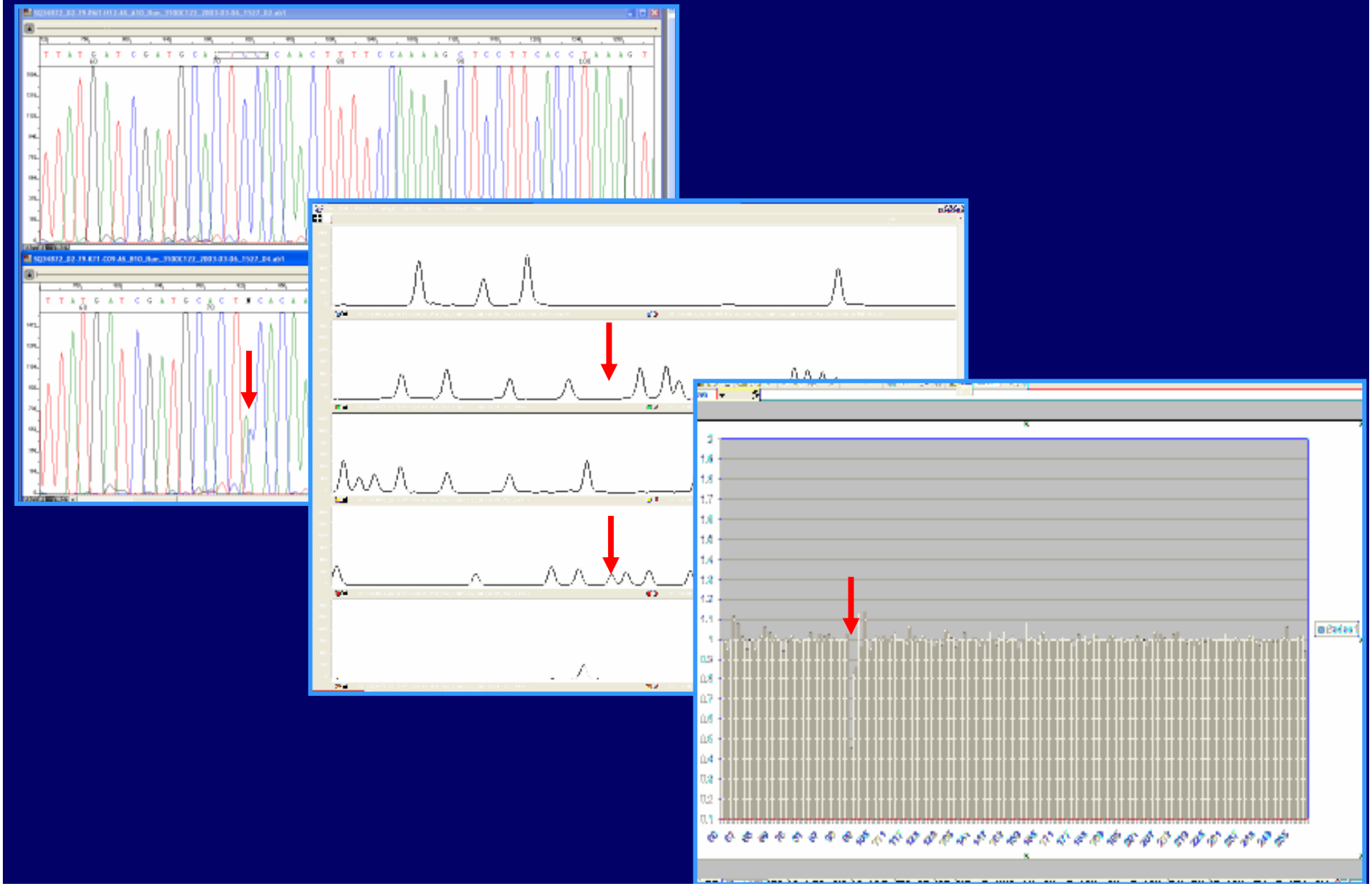
CSA



CSA



CSA



oncoCSA (Pearl/JAVA)

- Use Genescan/Genemapper to identify and quantify peaks.
- Use autoCSA to extract information for the real peaks and ignore noise peaks (using reference text)
- Compare normalised peak heights between sample and reference.
- If a normalised peak drops by >20%, ? Mutation (trace verify)

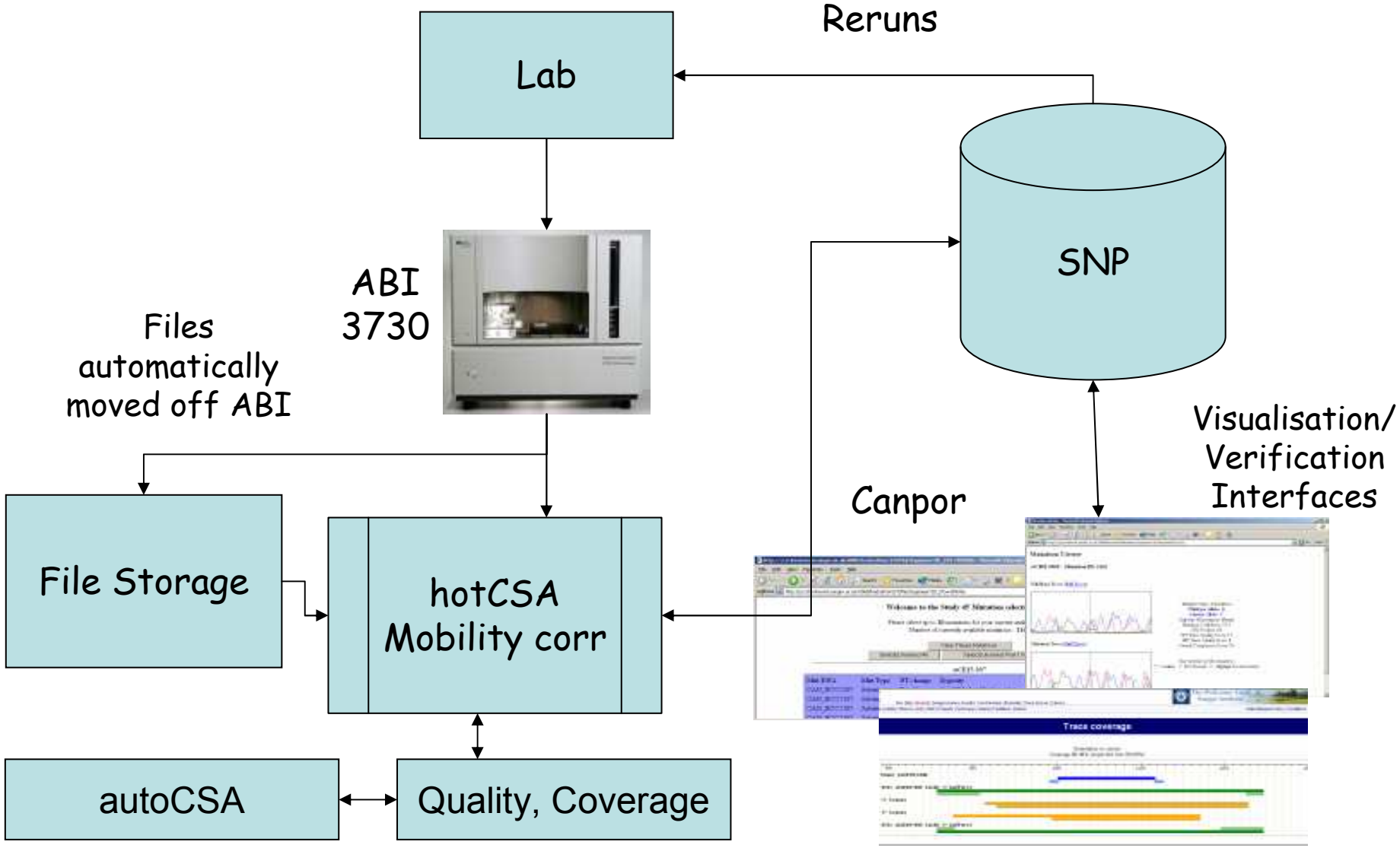
oncoCSA

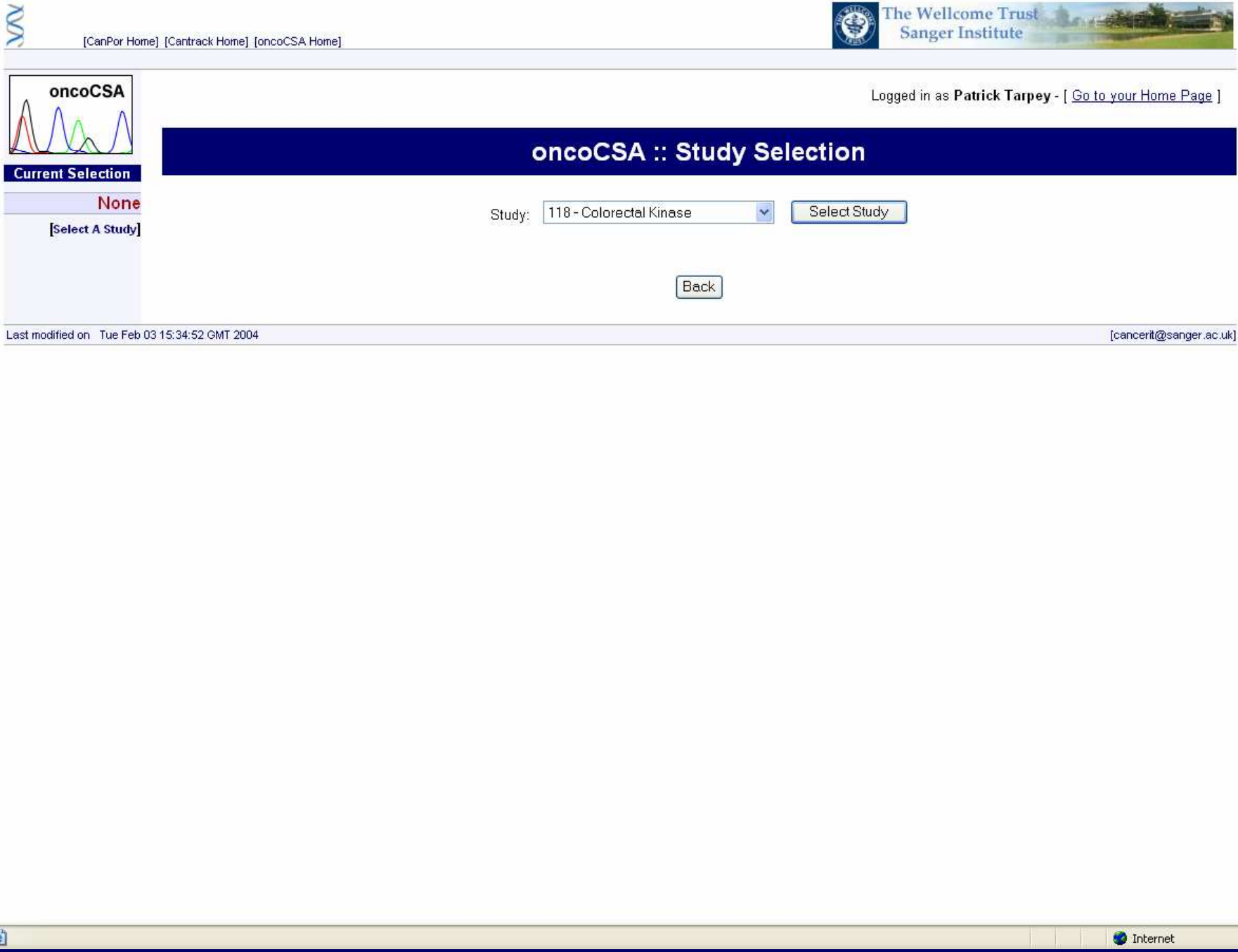
- Look for mutant peak (discriminate from noise)
- Present variant trace (and reverse)
- Apply a quality score for each base/trace (Q)
- Combine forward and reverse information

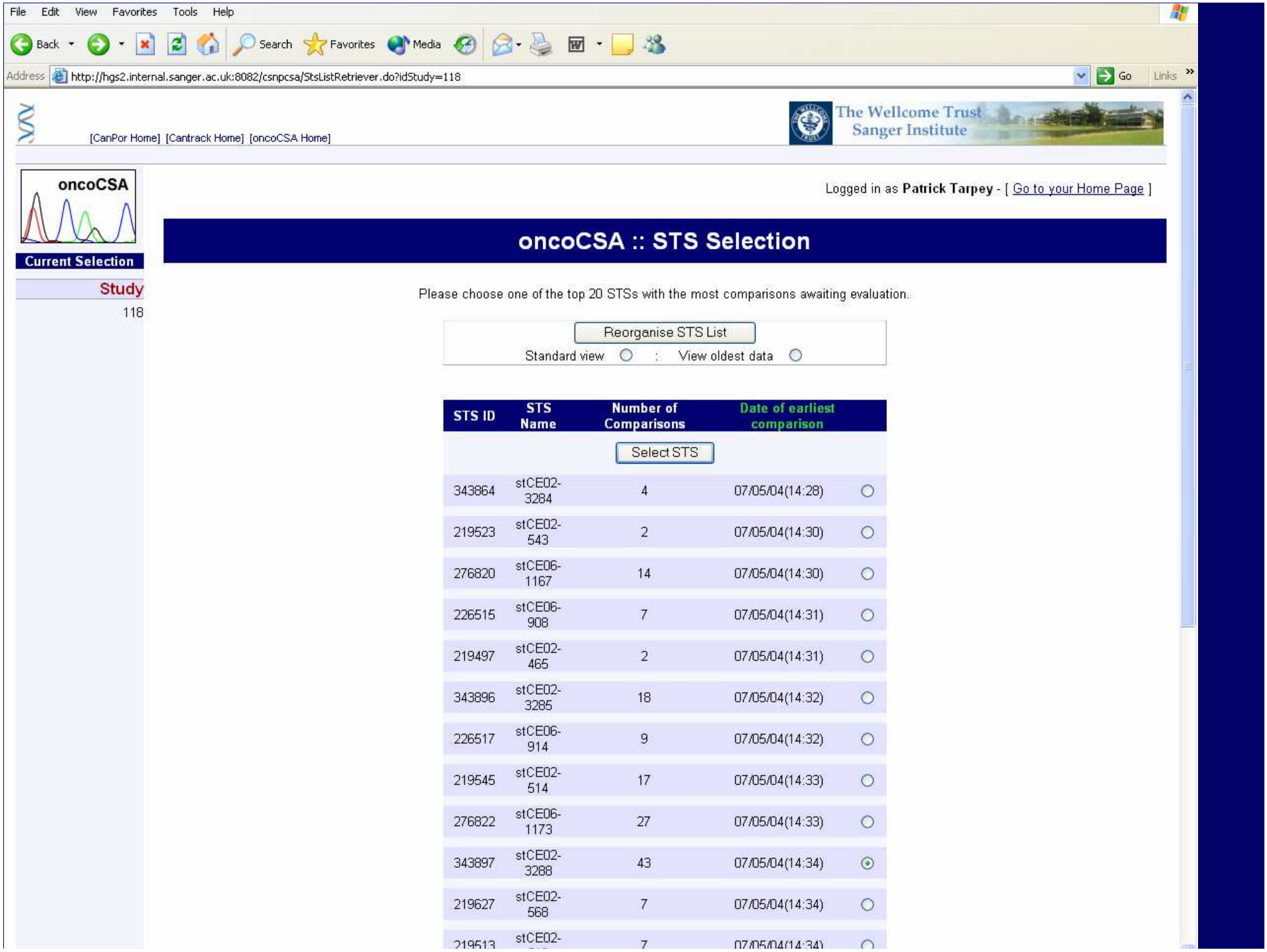
  (reduce false positives, determine coverage)

- Integrated Study management, tracking, analysis, annotation, archive

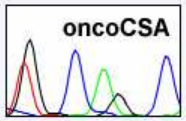
Analysis workflow




[CanPor Home] [Cantrack Home] [oncoCSA Home]
The Wellcome Trust Sanger Institute
oncoCSA
Logged in as Patrick Tarpey - [Go to your Home Page]
oncoCSA :: Study Selection
Current Selection
None
[Select A Study]
Study:
118 - Colorectal Kinase
Select Study
Back
Last modified on Tue Feb 03 15:34:52 GMT 2004
[cancerit@sanger.ac.uk]
Internet


File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://hgs2.internal.sanger.ac.uk:8082/csnpcsa/StsListRetriever.do?idStudy=118
Go Links
[CanPor Home] [Cantrack Home] [oncoCSA Home]
The Wellcome Trust Sanger Institute
oncoCSA
Current Selection
Study
118
Logged in as Patrick Tarpey - [Go to your Home Page]
oncoCSA :: STS Selection
Please choose one of the top 20 STSs with the most comparisons awaiting evaluation.
Reorganise STS List
Standard view : View oldest data
STS ID
STS Name
Number of Comparisons
Date of earliest comparison
Select STS
343864 stCE02-3284 4 07/05/04(14:28)
219523 stCE02-543 2 07/05/04(14:30)
276820 stCE06-1167 14 07/05/04(14:30)
226515 stCE06-908 7 07/05/04(14:31)
219497 stCE02-465 2 07/05/04(14:31)
343896 stCE02-3285 18 07/05/04(14:32)
226517 stCE06-914 9 07/05/04(14:32)
219545 stCE02-514 17 07/05/04(14:33)
276822 stCE06-1173 27 07/05/04(14:33)
343897 stCE02-3288 43 07/05/04(14:34)
219627 stCE02-568 7 07/05/04(14:34)
219513 stCE02- 7 07/05/04(14:34)

The Wellcome Trust Sanger Institute



Current Selection

Study

118

STS

stCE02-3288

Logged in as **Patrick Tarpey** - [Go to your Home Page]

oncoCSA :: Comparison Selection

Please select a **single** comparison for your current analysis needs

Current number of comparisons available: **43**

View Comparison Details & Mutations

Comparison ID	Mutant DNA (traceQual)	WildType DNA (traceQual)	Comp Score	Coverage Start	Coverage Stop	F/R	Comparison Date	Mutation Count	Select
98667	LS-1034 (15.82)	NCI-BL2052 (19.56)	42.0	20	202	r	May 7, 2004 2:34:32 PM	1	◉
98648	PD1505a (17.29)	HCC1937-BL (12.1)	70.0	68	242	f	May 7, 2004 2:34:26 PM	2	○
98649	PD1506a (17.38)	HCC1937-BL (12.1)	98.0	68	242	f	May 7, 2004 2:34:27 PM	1	○
98679	PD1506a (16.57)	NCI-BL2052 (19.56)	93.0	20	202	r	May 7, 2004 2:34:36 PM	1	○
98644	PD1507a (18.64)	HCC1937-BL (12.1)	84.0	68	242	f	May 7, 2004 2:34:25 PM	1	○
98647	PD1508a (16.74)	HCC1937-BL (12.1)	98.0	68	242	f	May 7, 2004 2:34:26 PM	1	○
98646	PD1509a (15.74)	HCC1937-BL (12.1)	58.0	68	242	f	May 7, 2004 2:34:26 PM	1	○
98645	PD1510a (15.38)	HCC1937-BL (12.1)	59.0	68	242	f	May 7, 2004 2:34:25 PM	1	○
98674	PD1510a (15.32)	NCI-BL2052 (19.56)	62.0	20	202	r	May 7, 2004 2:34:34 PM	1	○
98642	PD1511a (18.33)	HCC1937-BL (12.1)	38.0	68	242	f	May 7, 2004 2:34:25 PM	1	○
98672	PD1511a (6.29)	NCI-BL2052 (19.56)	69.0	20	203	r	May 7, 2004 2:34:34 PM	1	○
98643	PD1512a (16.28)	HCC1937-BL (12.1)	51.0	68	242	f	May 7, 2004 2:34:25 PM	1	○
98673	PD1512a (17.22)	NCI-BL2052 (19.56)	48.0	20	203	r	May 7, 2004 2:34:34 PM	1	○
98629	PD1513a (19.93)	HCC1937-BL (12.1)	93.0	68	242	f	May 7, 2004 2:34:20 PM	1	○

oncoCSA :: Mutation View

Current Selection

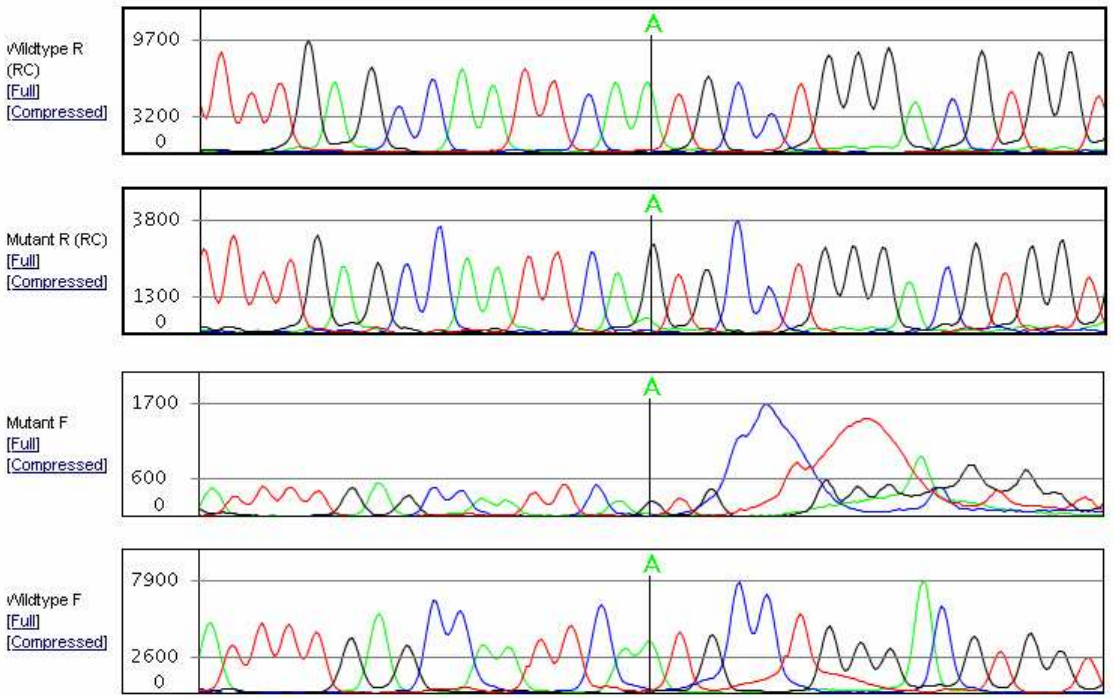
Study
118

STS
stCE02-3288

Comparison
NCI-BL2052(WT)
::
LS-1034(MT)

Mutations
406341

Substitution



Mutation 406341's Features	
Wildtype Allele	A
Mutant Allele	G
Amplimer	ATTCA A TGCCT
Type	Substitution
Zygosity	Homozygous Mutant
Confidence	12
Wildtype DNA	NCI-BL2052
Mutant DNA	LS-1034
STS Name	stCE02-3288
LoadingplateID	[LP994225]

[View Linked Traces]

Debug	
STS Position	98
Wildtype Trace Score	19
Mutant Trace Score	15
Comparison Score	42
Wildtype Peak Decrease	0.0
Sense Exon	true

Opinion

No opinion assigned to this mutant

Confirm	○
Reject	◉
Requires Second Review	○
None (leave till later)	○

[EDIT this Mutation] OR [ADD new mutation for this comparison]

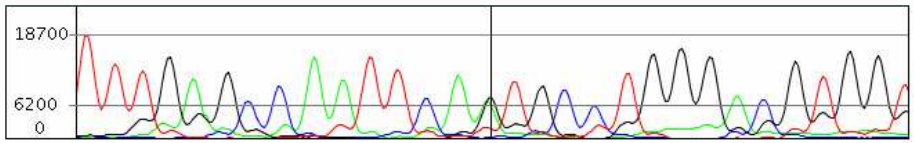
Trace Status

Fail First Trace (WT)	☐ (TraceID : 352238)
Fail Second Trace(MT)	☐ (TraceID : 422733)

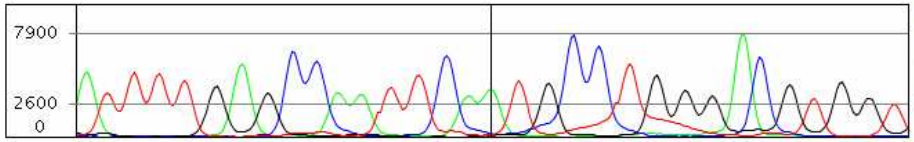
Done

Internet

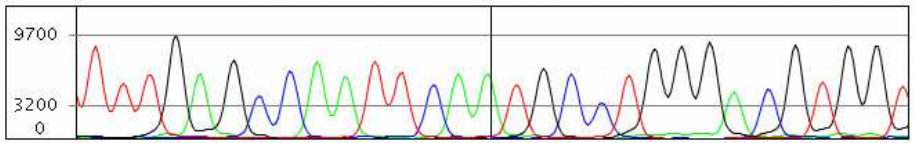
oncoCSA :: Parallel Trace View - Microsoft Internet Explorer



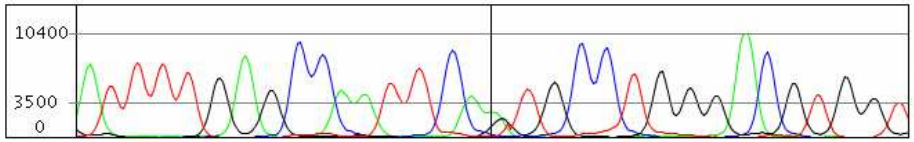
Trace: 322135
DNA : NCI-BL1770
Type : Normal
Direction : Reverse
Quality : 5



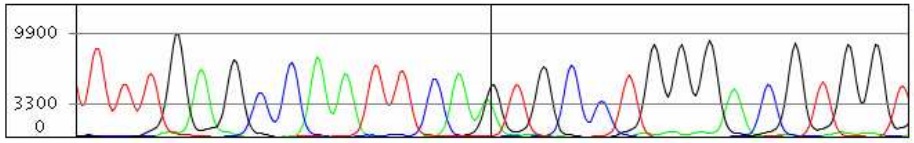
Trace: 352237
DNA : NCI-BL2052
Type : Normal
Direction : Forward
Quality : 19



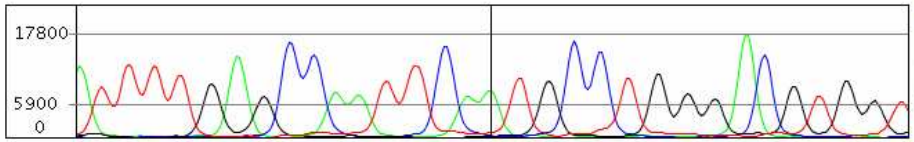
Trace: 352238
DNA : NCI-BL2052
Type : Normal
Direction : Reverse
Quality : 19



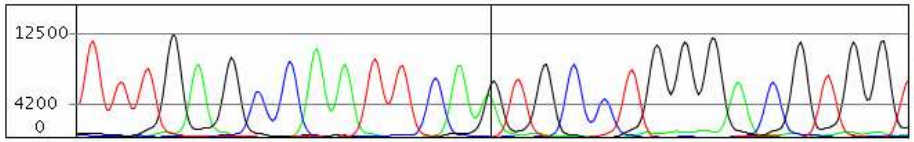
Trace: 352239
DNA : NCI-BL2171
Type : Normal
Direction : Forward
Quality : 20



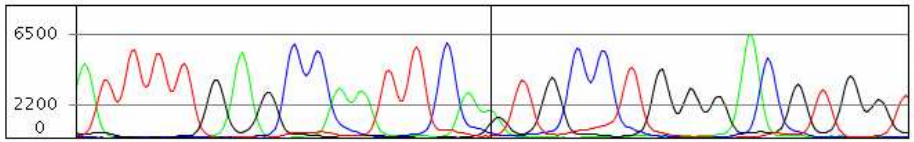
Trace: 352240
DNA : NCI-BL2171
Type : Normal
Direction : Reverse
Quality : 18



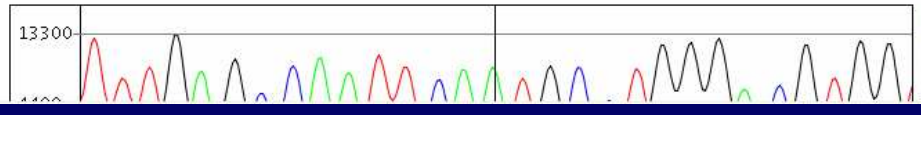
Trace: 352241
DNA : COLO-829-BL
Type : Mutant
Direction : Forward
Quality : 21



Trace: 352244
DNA : L542
Type : Mutant
Direction : Reverse
Quality : 17



Trace: 352245
DNA : L542
Type : Mutant
Direction : Forward
Quality : 17



Trace: 352246
DNA : COLO-829-BL
Type : Mutant

oncoCSA :: Mutation View

Current Selection

Study

11

ST

stCE02-328

Compariso

NCI-BL2052(W

LS-1034(M

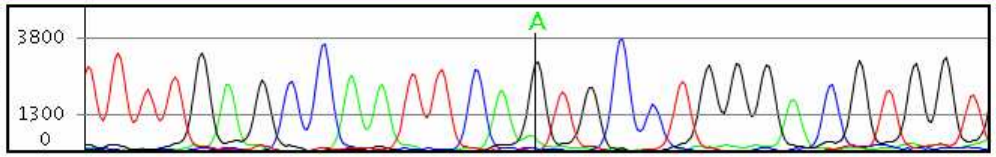
Mutation

40634

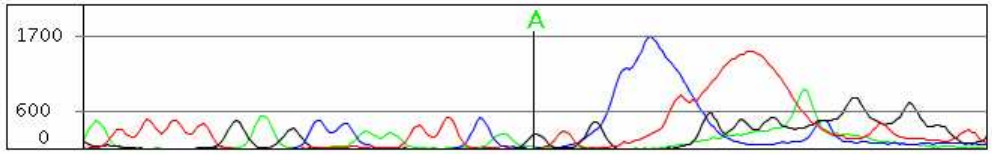
Full Trace View - TraceID: 352238 - Microsoft Internet Explorer

Center Trace | Close Trace

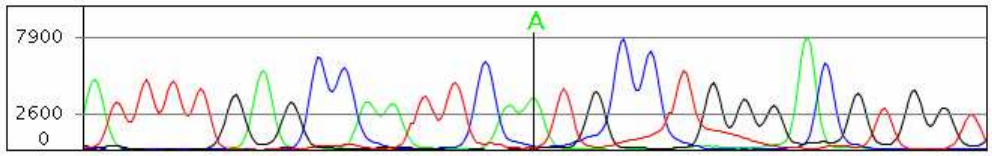
Mutant R (RC)
[Full]
[Compressed]



Mutant F
[Full]
[Compressed]



Wildtype F
[Full]
[Compressed]



Mutation 406341's Features

Wildtype Allele	A
Mutant Allele	G
Amplimer	ATTCA A TGCCT
Type	Substitution
Zygosity	Homozygous Mutant
Confidence	12
Wildtype DNA	NCI-BL2052
Mutant DNA	LS-1034
STS Name	stCE02-3288
LoadingplateID	[LP994225]

[View Linked Traces]

Debug

STS Position	98
Wildtype Trace Score	19
Mutant Trace Score	15
Comparison Score	42
Wildtype Peak Decrease	0.0
Sense Exon	true

Opinion

No opinion assigned to this mutant

Confirm	○
Reject	◉
Requires Second Review	○
None (leave till later)	○

[EDIT this Mutation] OR [ADD new mutation for this comparison]

Trace Status

Fail First Trace (WT)	☐ (TraceID : 352238)
Fail Second Trace(MT)	☐ (TraceID : 422733)

Done

Internet


[CanPan Home] [Cantrack Home] [oncoCSA Home]
The Wellcome Trust Sanger Institute
oncoCSA
Logged in as Patrick Tarpey - [Go to your Home Page]
Current Selection
Study
118
STS
stCE06-1167
oncoCSA :: Comparison Selection
Please select a single comparison for your current analysis needs
Current number of comparisons available: 14
View Comparison Details & Mutations
Comparison ID | Mutant DNA (traceQual) | WildType DNA (traceQual) | Comp Score | Coverage Start | Coverage Stop | F/R | Comparison Date | Mutation Count | Select
98040 | LS-1034 (18.48) | NCI-BL2171 (20.96) | 92.0 | 20 | 253 | r | May 7, 2004 2:30:59 PM | 1
98041 | PD1506a (5.13) | NCI-BL2171 (20.96) | 61.0 | 20 | 253 | r | May 7, 2004 2:31:00 PM | 1
98006 | PD1508a (3.15) | NCI-BL2171 (15.87) | 42.0 | 57 | 283 | r | May 7, 2004 2:30:48 PM | 2
98034 | PD1510a (16.76) | NCI-BL2171 (20.96) | 87.0 | 20 | 253 | r | May 7, 2004 2:30:58 PM | 1
98009 | PD1511a (3.95) | NCI-BL2171 (15.87) | 51.0 | 57 | 283 | r | May 7, 2004 2:30:49 PM | 2
98026 | PD1513a (4.72) | NCI-BL2171 (20.96) | 70.0 | 20 | 253 | r | May 7, 2004 2:30:55 PM | 2
98029 | PD1517a (19.52) | NCI-BL2171 (20.96) | 82.0 | 20 | 253 | r | May 7, 2004 2:30:56 PM | 1
98031 | PD1520a (4.4) | NCI-BL2171 (20.96) | 68.0 | 20 | 253 | r | May 7, 2004 2:30:57 PM | 2
98019 | PD1521a (19.77) | NCI-BL2171 (20.96) | 79.0 | 20 | 253 | r | May 7, 2004 2:30:53 PM | 1
98018 | PD1522a (19.63) | NCI-BL2171 (20.96) | 96.0 | 20 | 253 | r | May 7, 2004 2:30:53 PM | 1
98021 | PD1524a (19.79) | NCI-BL2171 (20.96) | 58.0 | 20 | 253 | r | May 7, 2004 2:30:54 PM | 1
97995 | PD1525a (3.46) | NCI-BL2171 (15.87) | 72.0 | 58 | 283 | r | May 7, 2004 2:30:45 PM | 1
98024 | PD1528a (19.21) | NCI-BL2171 (20.96) | 64.0 | 20 | 253 | r | May 7, 2004 2:30:55 PM | 1
98046 | PD1531a (18.5) | NCI-BL2171 (20.96) | 62.0 | 20 | 253 | r | May 7, 2004 2:31:01 PM | 1
Internet

[CanPor Home] [Cantrack Home] [oncoCSA Home]

The Wellcome Trust Sanger Institute

oncoCSA

Logged in as **Patrick Tarpey** - [Go to your Home Page]

oncoCSA :: Comparison Selection

Current Selection

Study
118

STS
stCE06-1167

Comparison ID	Mu
98040	LS
98041	PD
98005	PD
98034	PD
98009	PD
98026	PD
98029	PD
98031	PD
98019	PD
98018	PD
98021	PD
97995	PD
98024	PD
98046	PD

[CanPor Home] [Cantrack Home] [oncoCSA Home]

The Wellcome Trust Sanger Institute

oncoCSA

Logged in as **Patrick Tarpey** - [Go to your Home Page]

oncoCSA :: Mutation View

Current Selection

Study
118

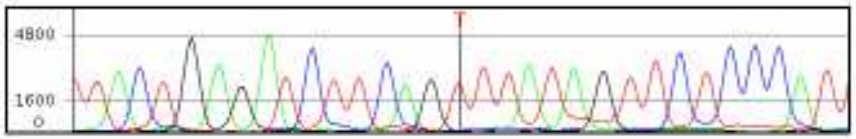
STS
stCE06-1167

Comparison
NCI-BL2171(WT)
PD1506a(MT)

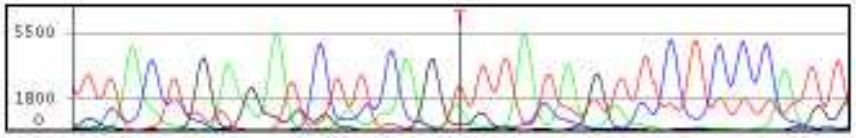
Mutations
405413

Substitution

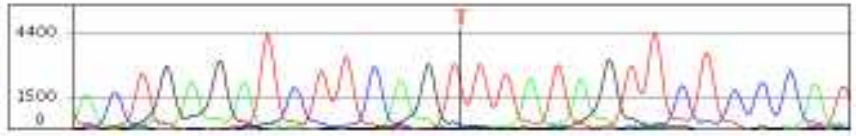
Wildtype R [Full] [Compressed]



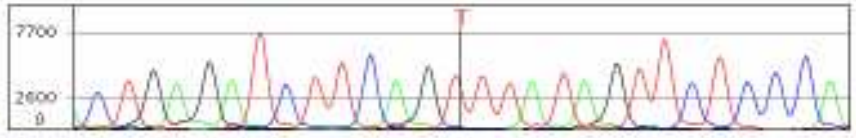
Mutant R [Full] [Compressed]



Mutant F (RC) [Full] [Compressed]



Wildtype F (RC) [Full] [Compressed]



Mutation 405413's Features

Wildtype Allele	T
Mutant Allele	A
Amplimer	TTCAG T TTATA
Type	Substitution
Zygosity	AutoCSA Amplification value
Confidence	15
Wildtype DNA	NCI-BL2171
Mutant DNA	PD1506a
STS Name	stCE06-1167
LoadingplateID	[LP0062268]

[View Linked Traces]

Debug

STS Position	98
Wildtype Trace Score	20
Mutant Trace Score	5
Comparison Score	81
Wildtype Peak Decrease	0.744
Sense Exon	false

Opinion

No opinion assigned to this mutant

Confirm ○
Reject ◉
Requires Second Review ○
None (leave till later) ○

[EDIT this Mutation] Off [ADD new

Done

Internet

Validation

Competition set

- Sensitivity of Substitution detection
  - Heterozygotes = 98% (4 not interogated, missed 2)
  - Homozygotes = 100%
  - (data set of 7,733 comparisons compared against Mutation surveyor)

  - CSA detected 3 additional mutations Surveyor missed

oncoCSA SUMMARY

CSA development

- CSA looks robust at high-throughput
- Integrated sample management, tracking, analysis, annotation, archive

Current development

- Q scores maximise information retrieved from a trace
- het insdels

Surveyor

- Good for het point mutations
- Not 100% sensitive
- Controls
- BLAST?

Cancer Genome Project

Mike Stratton

Andy Futreal

Richard Wooster

Bamford, Sally,
Barthorpe, Andrew
Bignell, Graham
Blow, Matthew
Brackenbury, Lisa
Butler, Adam
Chim, Anne
Clarke, Oliver
Clegg, Sheila
Cole, Jennifer
Davies, Helen
Dawson, Elisabeth
Dicks, Ed
Dike, Angus
Drozd, Anja
Edkins, Sarah
Edwards, Ken
Forbes, Simon
Foster, Rebecca
Gray, Kristian
Greenman, Chris
Halliday, Kelly
Haynes, Wendy
Hills, Katy
Korny, Angelique
Kosmidou, Vivienne
Lugg, Richard
Menzies, Andrew
O'Meara, Sarah
Parker, Adrian
Perry, Janet
Petty, Robert
Raine, Keiran
Ratford, Lewis
Shepherd, Rebecca
Small, Alexandra
Smith Rafaella
Stephens, Philip
Stephens, Yvonne
Stevens, Claire
Tarpey, Patrick
Teague, Jon
Tofts, Calli
Varian, Jennifer
West, Sofie
Widaa, Sara
Yates, Andrew

X project

Lucy Raymond
Josep Parnau

UK, Aus, USA, Europe Clinicians

CSA, Cambridge, Salisbury

Chris Mattocks

Jo Whittaker

Martin Bobrow