GeneScreen: a program for high-throughput mutation detection in DNA sequence electropherograms

Ian M Carr, Nick Camm, Graham R Taylor, Ruth Charlton, Sian Ellard, Eamonn G Sheridan, Alexander F Markham, David T Bonthron

GeneScreen: a program for high-throughput mutation detection in DNA sequence electropherograms

Ian M. Carr,[1*] Nick Camm,[2] Graham R. Taylor,[1,2] Ruth Charlton,[2] Sian Ellard,[3] Eamonn G. Sheridan,[1,2] Alexander F. Markham,[1] David T. Bonthron[1]

[1]Leeds Institute of Molecular Medicine, University of Leeds, UK; [2]Yorkshire Regional Genetics Service, St. James's University Hospital, Leeds, UK; [3]Department of Molecular Genetics, Royal Devon & Exeter NHS Foundation Trust, Exeter, UK

*Correspondence to:

Ian M. Carr
Leeds Institute for Molecular Medicine
Level 9, Wellcome Trust Brenner Building
St. James's University Hospital
Leeds LS9 7TF, U.K.

E-mail: i.m.carr@leeds.ac.uk

Keywords: mutation detection, software, DNA sequencing

Word count: 3729

ABSTRACT

Background

While massively parallel DNA sequencing methods continue to evolve rapidly, the benchmark technique for detection and verification of rare (particularly disease-causing) sequence variants remains four-colour dye-terminator sequencing by capillary electrophoresis. The high throughput and long read lengths currently available have shifted the bottleneck in mutation detection away from data generation to data analysis. While excellent computational methods have been developed for quantifying sequence accuracy and detecting variants, either during *de novo* sequence assembly or for SNP detection, the identification, verification and annotation of very rare sequence variants remains a rather labour-intensive process for which few software aids exist.

Aim

To provide a freely available, intuitive software application for highly efficient mutation screening of large sequence batches.

Methods and Results

We developed *GeneScreen*, a desktop program that analyses capillary electropherograms and compares their sequences to a known reference for identification of mutations. The detected sequence variants are then made available for rapid assessment and annotation via a graphical user interface, allowing chosen variants to be exported for reporting and archiving. The program was validated using more than 16,000 diagnostic laboratory sequence traces.

Conclusion

Using *GeneScreen*, a single user requires only a few minutes to identify rare mutations in hundreds of sequence traces, with comparable sensitivity to expensive commercial products.

INTRODUCTION

DNA sequence analysis is a central technology in many branches of biology, perhaps most critically where the relationship between genotype and phenotype is under scrutiny. The massive expansion in the catalogue of human disease-associated genetic variation over the past three decades has been parallelled by the development of diverse methods for detection of sequence variants, particularly approaches that are easily scalable for screening large numbers of DNA samples. Landmark techniques for detecting unknown sequence variants (typically in PCR-generated genomic templates) include chemical heteroduplex cleavage[1], denaturing gradient gel electrophoresis[2 3], single-stranded conformation polymorphism[4] and high-resolution melting curve analysis[5]. The driver for the development of most such methods has been the labour-intensiveness of complete DNA sequencing (earlier by manual and more recently by semi-automated versions of the Sanger approach). Although some of these mutation screening methods can offer high sensitivity for detection of unknown variants, this may demand careful optimization of experimental conditions to suit individual amplicons. Consequently, as modern versions of the Sanger sequencing method have improved in speed and quality[6 7], sequencing has persisted as the commonest method of choice for mutation detection. Its near-universal applicability under standard conditions makes it particularly attractive for laboratories analysing a wide range of rare single-gene disorders (for whom the optimization of other methods for analysis of diverse genomic targets can present major challenges). Despite the current rapid development of clonal sequencing technologies, high-quality Sanger sequencing of medium-sized (100–1000 bp) PCR-amplified templates is likely to remain a benchmark technology for mutation detection and verification for the foreseeable future.

A major strength of Sanger sequencing is that it provides information on each nucleotide sequenced over a read length of up to 1 kb. Currently, its main limitation in many laboratories is that the large quantity of sequence data generated by modern instruments is both tedious and time-consuming to search for variant positions. Analysis of electropherograms is also highly prone to human error. For large sequencing projects automated analysis of sequence electropherograms has been an absolute requirement. Techniques arising out of the Human Genome Project tended to

focus on the detection of sequence variants (including sequencing errors) in sequences of cloned DNA, with the high quality data then used to construct sequence contigs[8 9]. This high quality data allows detection of polymorphisms through the presence of variant nucleotides at the same position in multiple traces[10]. Since individual sequences originate from clones, this approach does not require computer programs to deal with heterozygous trace positions (which in this setting would indicate sequence artefact, register a low quality score and so be omitted from subsequent analysis). However, as the focus widened to include SNP detection and resequencing projects, software was developed that could genotype heterozygous positions present in PCR products.[10 11] Such applications identify heterozygous positions by the presence of two overlapping peaks of similar height or by a direct comparison to a reference trace file[12]. The presence of a SNP can be verified by the identification within the study population of samples homozygous for either allele.

While these approaches work well for the detection of SNPs, additional problems complicate the discovery of rare pathological variants. For example, for *de novo* disease-causing mutations, and even for typical recessive diseases that show extreme allelic heterogeneity, it is unlikely to be possible to observe a second example of a heterozygous mutation among a cohort of samples; homozygosity for the rare allele will also not be observable. Consequently, there will be no supporting evidence to validate the variant within the study cohort. Also, while many heterozygous positions show two overlapping electropherogram peaks of similar heights, this is not always true; heterozygous positions where the allelic peaks have different migration rates and/or peak heights are quite common. When screening for SNPs to describe genetic variability within a population, some level of false negative results is acceptable. In contrast, when searching for pathological changes, it is more important to minimize the false negative rate, even at the cost of a false positive rate requiring assessment by a human operator.

Consequently, while it is possible to create a highly automated process for SNP detection, most diagnostic and research laboratories searching for rare pathogenic variants have not felt able to rely entirely on computer programs to automate their detection. We have therefore focussed our attention on maximizing the efficiency with which a human operator, with the aid of the computer, can detect mutations within a large number of electropherograms and assess their likely significance. The

application we have developed, *GeneScreen*, automates the process of detecting possible sequence variants and then, critically, presents the data in a manner that allows a user very quickly to categorize and annotate variants as true or false positives. The program then allows the easy export of the annotated variants for reporting and cataloguing. *GeneScreen*'s focus on a highly efficient user interface, and its integrated capability for annotating rare variants in standard nomenclature on genomic, cDNA or protein reference sequences, are points of difference from previous applications, such as inSNP[13], varDetect[14] and novoSNP[15], all of which offer various capabilities for mutation detection and annotation, but have not been widely adopted by genetic diagnostic laboratories.

MATERIALS AND METHODS

Programming

Programming was done using Microsoft Visual Studio 2005. *GeneScreen* runs on Microsoft Windows operating systems, and requires the .NET framework 2.0. Freely available downloads and detailed user guides are at http://dna.leeds.ac.uk/genescreen.

Sequencing data

Sequence electropherograms were generated on Applied Biosystems (Foster City, CA, USA) capillary sequencing instruments in the NHS Regional DNA Laboratories in Leeds and Exeter and the CRUK Genomic Services Mutation Detection Facility (Leeds), either as part of internal quality control of diagnostic sequencing, or for external research projects. (For *ABCC8*, the patients and sequences were a subset of those previously described[16].) Only sequence variants in exons or within 20 bp of a splice site were annotated. According to the priorities of each sequencing project, only selected exons were sequenced in some genes, while in others, larger exons were analysed as a number of overlapping sequences. For example, only protein-coding exons were analysed; the alternatively spliced exon 4 from *BRCA1* was not sequenced; *BRCA1* exon 10 and *BRCA2* exons 10, 11 and 27 were amplified as several shorter products. The *TP53* sequences were derived from archived tumour samples, while the *ABCC8*, *BRCA1*, *BRCA2* and *GCK* DNA were from patients suspected of having germline mutations, based on the presentation of disease either in the subject under analysis or in a close family member.

For a random subset of the *ABCC8*, *GCK* and *TP53* sequence files, the ABI Sequencing Analysis v5.2 program was used to create phd.1 files (containing Phred-like quality scores). These files were then processed using the program "QVregion.exe" (http://dna.leeds.ac.uk/qvregion/) in order to obtain the quality scores for the ten worst nucleotides, over the region screened for sequence variants (as defined above—exons + 20 bp of flanking intron).

Data analysis workflow

In outline, *GeneScreen* detects sequence variants by following a multistep process for each ABI file. Initially the fluorescence data from each ABI file is extracted and used to deduce the nucleotide sequence. This is then aligned to the reference sequence enabling any sequence variants to be identified. Finally, any sequence variants identified are annotated at the genomic, cDNA and protein sequence level where appropriate. It is important to note that Genescreen base-calls each trace independently of the reference or other sequences in the dataset and does not attempt to correct sequencing artefacts. The data analysis workflow is described in greater detail in the Supplementary Data.

RESULTS

Method of use

A full illustrated description of *GeneScreen* usage is at http://dna.leeds.ac.uk/genescreen/guide/. A brief outline of certain aspects follows.

The initial tabbed user interface of *GeneScreen* (**Figure 1**) allows access to each of several sequential steps required for scrutiny of a large batch of test sequences.

Reference file creation

To identify the positions of sequence variants, *GeneScreen* requires a reference file. This can be either a plain text sequence, or a special gene descriptor reference file created by *GeneScreen* itself, using as input cDNA and genomic sequences plus an optional tab-delimited list of the names and sequences of the sequencing primers. The gene descriptor file specifies the position of each exon, relative to the start codon and the beginning of the genomic sequence. This information is then used by *GeneScreen*

to annotate sequence variants according to Human Genome Variation Society nomenclature (http://www.hgvs.org/index.html).

Identification and display of sequence variants

Once a reference file is available, a directory containing a large batch of electropherograms can be loaded and analyzed (taking ~15 s per 96 traces aligned to a 250-bp reference sequence on a typical 2.4 GHz desktop processor). *GeneScreen* can be configured to reject sequences containing more than a pre-specified number of sequence variants, and the minimum peak height ratio for a heterozygous call is also adjustable. *GeneScreen* then displays the collated information on all detected sequence variants in the form of a data grid (**Figure 2**), within which each column represents a position at which a sequence variant exists in one or more trace files, while each row corresponds to one electropherogram. Provided appropriately annotated reference files were used, positions within exons are coloured blue, and introns white. Locations where a sequence deviates from the reference sequence appear as red cells, with the nucleotide substitution in black text. For each cell it is possible to view its underlying sequence data as an image of the electropherogram around the mutation, or a comparison of the mutant electropherogram to a control trace with normal sequence. The grid and displayed image can be dynamically linked, so that simply moving the mouse cursor across the grid displays the local sequence trace corresponding to the underlying cell, allowing very rapid visual inspection (Figure 2). Left-clicking on a red cell invokes a floating menu that allows the user to annotate and save the sequence variant, to genotype SNPs, or to analyse and annotate complex mutations. Even for heterozygous insertions or deletions, *GeneScreen* can usually perform a correct subtractive analysis of the superimposed sequences downstream of the mutation point, correctly calling and annotating the mutation, as shown for the examples below in **Figure 3**. All the analysis and annotation options are shown in detail in the online program guide.

Exporting and archiving sequence variants

Selected sequence variants can be saved as a tab-delimited plain text (with or without linked images of the variant), a LOVD-compatible data import file, or as a web page that includes images of the exported sequence variants (**Figure 3**). (Note that *GeneScreen*'s LOVD-compatible file is designed to match standard LOVD2

installations.) If a different layout is required, the form can easily be edited using a spreadsheet application. To extend the utility of the web page output format, new variants can be incrementally added to an existing page, such that all the variants are ordered by gene name, position in the genomic reference sequence and patient ID (irrespective of the order in which they were originally selected for export). *GeneScreen* also has a web page editing function that allows the variants in the web page to be edited, updated or deleted. As a result of these features, the web page serves the dual functions of a data presentation file and a database of selected sequence variants.

Design decisions: sequencing artefacts and miscalls

GeneScreen is designed for rapid interactive analysis of large sequence batches by a human operator, not for full automation. It leaves the operator to make decisions related to common sequencing problems, such as "dye blobs", low signal, high background and anomalous peak mobility, and does not introduce computational corrections for such anomalies. This decision stems from the desire to avoid computational artefacts, including the overlooking of genuine sequence variants. Some illustrations of these concerns are shown in Figure 4. Firstly, **4A and 4B** show forward and reverse sequences around a heterozygous substitution. In 4A, the different mobilities of the allelic peaks have resulted in a miscalled insertion. A computational correction for this mobility shift could be applied, but this might adversely affect other traces, such as that in **Figure 4C**; here, software "correction" of the overlapping C and A peaks would result in an incorrect heterozygous call. (The reverse sequence in Figure 4D reveals the peak overlap to be another artefact of anomalous migration.)

Resolving doubtful cases by reference to the reverse strand sequence, as in both of these examples, is not always possible, for various reasons. Firstly, it may not be obvious which of two discordant forward and reverse sequences has been miscalled. Secondly, forward and reverse sequences may not both be available, either for reasons of economy or technical difficulty (as when sequencing through heterozygous indels, or long simple tandem repeats). Finally, forward and reverse sequences may occasionally be wrongly paired through operator error, as a result of which true

heterozygous variants might be discarded, because of appearing to be absent from the sequence in the other orientation.

Correcting for variation in peak height can also result in conflicting results. **Figure 4E** shows a heterozygous C/T call at a position which in the reverse sequence (**Figure 4F**) is clearly homozygous C. This error results from the weakness of the genuine C peak in Figure 4E, compared both to flanking positions and to the background. If a scaling correction (based on a reference trace) were used to accept the small blue C peak as genuine, while discarding the red T signal as background, a risk would be introduced of miscalling similarly appearing positions where two small superimposed peaks do in fact reflect true heterozygosity.

Similar to Figure 4E, **Figure 4G** shows a miscall due to a sequencing artefact that lies between two peaks. The artefactual peak is of comparable height to the genuine C in Figure 4E. It is therefore apparent that any filtering aimed at disregarding the artefactual peak in Figure 4G might adversely affect the calling of traces similar to Figure 4E.

Finally, using peak height ratios as a criterion of heterozygosity can be problematic. **Figure 4I-J** shows a position that is clearly heterozygous in the reverse sequence, but for which the allelic peak heights differ greatly in the forward sequence. A base-caller configured to use peak height ratios to ignore the spurious peak in 4E might also be liable to miscall traces such as 4I.

In most situations, overlooking a true sequence variant has more serious implications than identification of false positives. *GeneScreen* has therefore been designed with default settings and a user interface that minimize the impact of dealing with a large number of false positive calls. Rapid real-time visual inspection of all suspect positions within a large number of traces (including comparison of paired forward and reverse reads) allows correct decisions to be made with minimal user intervention.

Analysis of test sequences

To test *GeneScreen*'s accuracy in detecting sequence variations, we used it to analyse several thousand sequences from *ABCC8*, *BRCA1*, *BRCA2*, *GCK* and *TP53* (**Table 1**). Each set of test data was analysed using both *Mutation Surveyor* (SoftGenetics, State College, PA) and *GeneScreen*, by operators familiar with each piece of software.

Each program detected an identical set of 177 rare sequence variants (Table 1). We also found *GeneScreen* to be at least as fast as *Mutation Surveyor*. (For example, the 30 amplicons covering the *BRCA1* open reading frame of 96 subjects were analysed by one operator in ~5 hours, compared to ~9 hours using *Mutation Surveyor*.) *GeneScreen* also correctly genotyped all the common variants (SNPs) present in the test sequence files. It should be noted, though, that the identification of common SNPs is rather easy for a mutation-screening program to perform, because they show up as sequence variants at the same position in multiple sequence traces. They are therefore excluded from Table 1. Apart from common SNPs, none of the sequence variants in *ABCC8*, *BRCA1*, *BRCA2* or *GCK* was homozygous. Most of the *TP53* variants likewise appeared heterozygous, but in some cases LOH in tumour DNA resulted in variants appearing to be homozygous.

Gene	**Subjects**	**Exons**	**Amplicons***	**Traces**	**Substitutions**	**Indels**
ABCC8	31	39	39	1209†	20	2
BRCA1	96	24	30	5760	15	3
BRCA2	96	27	32	6144	5	6
GCK	160	10	10	1600†	49	5
TP53	95	11	10	1900	57	15
Total				**16613**	**146**	**31**

Table 1

Description of test sequences used to validate *GeneScreen* and mutations identified (excluding common SNPs). *See main text for discrepancy between number of exons and number of PCR products. †Amplicons were only sequenced in the forward direction. “Indels” refers collectively to deletions, insertions and insertion-deletions.

Sequence quality and miscall rate

To assess whether the false positive variants identified by *GeneScreen* resulted in most part from poor quality sequence, or from poor performance of its base-caller, we

analysed a random subset of our traces using the commercial ABI Sequencing Analysis v5.2 program, which assigns per-residue quality values defined in similar fashion to Phred qualities[9]. (We reasoned that if the false positives were attributable to a weak base-caller, quality values of reads containing false positives would be similar to those of files containing no variants at all.) Additionally, because sequence artefacts resulting in false positives might reflect either globally poor signal-to-noise or discrete problems such as 'dye blobs' or compressions, we used an approach of categorizing sequence quality according to the average quality values of the ten worst nucleotides in a sequence. This was found empirically to allow detection of both generally poor quality and short regions of poor sequence in otherwise good reads, while not being overly sensitive to the presence of true heterozygous variants (which also lower the quality score). **Figure 5** shows that traces in which *GeneScreen* identified a false positive variant tend to be those with the lowest quality values as measured in this way.

Of the 2143 trace files used for this analysis, 152 (6.8%) generated at least one false positive variant call, while 123 (5.4%) contained at least one true variant (including SNPs). Predictably, the latter group tended to contain only one variant per read (133 variants in 123 sequences), in contrast to reads with false positives (328 variants in 152 sequences). Overall, 29% of all variants identified in this analysis were true variants. Within the 385 628 nt of sequence coverage, this corresponded to a false positive rate of 1 per 1175 nt (compared to a true positive rate of 1 per 2899 nt).

We also noted a strongly non-random distribution of the false positive results, according both to the identity of the individual PCR product (*i.e.* its sequence) and to the quality of the sequencing template (depending in turn on genomic DNA quality). For this reason, we have not attempted to derive statistical measures of the sensitivity and specificity of *GeneScreen*; experimental variation, or choice of different genomic targets, would render such figures invalid. For example, our *TP53* DNA samples had been extracted from archived tumour samples, and consistently produced higher false positive rates than samples extracted from blood. Similarly, sequences that involved reading through a long mononucleotide run were more susceptible to false positive results.

DISCUSSION

Identification of unknown functionally important rare sequence variants is an important activity in many research and diagnostic laboratories. The accuracy and volume of sequence data available at a given cost have steadily increased over the last decade, exposing the less standardized process of data analysis as increasingly tedious and error-prone. It is not surprising, therefore, that a number of computer programs have been developed with the aim of screening electropherograms for the presence of mutations. The fact that most diagnostic DNA laboratories continue to inspect patient sequences visually presumably reflects perceived limitations of available software. Aside from the question of sensitivity, these perceived limitations may include aspects such as complexity of the software, platform-dependence, or its targetting primarily at other uses such as *de novo* sequence assembly or SNP detection. Unlike large genome centres (that may be focussed on the latter areas) diagnostic and small genetics research laboratories have generally not had the expertise to undertake dedicated software development. Ease of use is one reason for the wide popularity in diagnostic laboratories of the commercial *Mutation Surveyor* package (SoftGenetics LLC, PA, USA).

Programs aimed mostly at identifying SNPs can improve their specificity using population criteria, such as a requirement for each allele of a putative variant to occur a certain number of times in a set of sequence traces. This approach to reducing the false positive background is not applicable to rare disease-causing variants. Developed primarily with the latter goal, *GeneScreen* therefore base-calls each sequence independently and then flags variants after comparison to the known reference sequence. No attempt is made to detect and remove false positives by reference to external data, since in some circumstances this would be likely to result in the overlooking of true positives that appear similar to sequencing artefacts (Figure 4). The authors strongly believe that the best way to keep the number of false positives manageable is to produce high quality sequence data.

To compensate for the variability in peak height and mobility that are intrinsic to dye-terminator sequencing, some mutation detection programs employ direct subtractive analysis of the difference between test and control sequence traces. As well as *Mutation Surveyor*, this approach is used by the *tracediff* component of the Staden

package[17] and *SeqDoC* [12]. This method can perform excellently when sequence batches derive from templates of highly reproducible quality and quantity. It can be rather susceptible, though, to differences in quality and signal intensity between test and control samples[12]. Since such sample variability is common and often unavoidable when dealing with patient material, we decided against such an approach in the design of *GeneScreen*. A brief summary of the features of a number of mutation detection programs is shown in **Table 2**.

Program name	Variant detection method (i)	Variant annotation (ii)	User friendly (iii)	Freeware	Analyses each file independently
Genescreen	Type I	G, C and P	Yes	Yes	Yes
InSNP[13]	Type I	R	Yes	Yes	Yes
Mutation Surveyor	Type II	G, C and P	Yes	No	Yes
novoSNP[15]	Type I	R	Yes	Yes	Compares variants in forward and reverse reads when possible.
Polyphred[11 18]	Type I	R	No	Yes	Compares variants in forward and reverse reads when possible.
SeqDoc[12]	Type II	R	Yes	Yes	Yes
SNPDetector[19]	Type I	R	Yes	Yes	No
Staden package[17]	Type II	G, C and P	No	Yes	Yes
VarDetect[14]	Type I	R	Yes	Yes	Yes

Table 2

Feaures of some DNA sequence variant detection programs. (i) The programs detect variants by either by base-calling the trace files and aligning the sequence to the reference file (Type I) or by direct subtractive analysis between test and control traces (Type II). (ii) Variants are annotated relative to genomic sequence (G), cDNA sequence (C), protein sequence (P) or an arbitrary reference sequence (R). (iii) Subjectively refers to ease of setup and "learning curve" for those who are not computer specialists.

Platform-dependence is a perennial issue that has undoubtedly affected the adoption by less technically minded users of some highly sophisticated software packages (such as the Staden package, originally developed solely for a Unix/Linux environment). Limited user acceptance can also result from lack of familiarity with the command line environment through which some programs such as *PolyPhred* [11] are controlled. However, development of a truly cross-platform application inevitably involves greater effort than focussing on one operating system. To keep development effort manageable, *GeneScreen* utilizes the .NET framework and is hence not cross-platform. This choice, though, does ensure that it is accessible to most non-specialist computer users.

Compared to some other programs for sequence variant detection, *GeneScreen* does not go to great lengths to maximize the specificity with which variant positions are identified; this is in order to avoid loss of sensitivity for detection of rare variants. Our design consequently places considerable onus on the operator to distinguish true from false positives, and to that end, the grid presentation of *GeneScreen* maximizes the efficiency with which the user can inspect possible variant positions. The higher the sequence quality, the fewer artefactual variants are identified on the grid, and the quicker the task of analysis; consequently efficient mutation identification requires high quality sequence.

REFERENCES

1. Dahl HM, Lamande SR, Cotton RGH, Bateman JF. 1989. Detection and localization of base changes in RNA using a chemical cleavage method. Anal Biochem 183:263–268.

2. Fischer SG, Lerman LS. 1983. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. Proc Natl Acad Sci USA 80:1579–1583.

3. Sheffield VC, Cox DR, Lerman LS, Myers RM. 1989. Attachment of a 40-base-pair G + C-rich sequence (GC-clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single-base changes. Proc Natl Acad Sci USA 86:232–236.

4. Boothroyd CV, Teh BT, Hayward NK, Hickman PE, Ward GJ, Cameron DP. 1991. Single base mutation in the hormone binding domain of the thyroid hormone receptor beta gene in generalised thyroid hormone resistance demonstrated by single stranded conformation polymorphism analysis. Biochem Biophys Res Comm 178:606–612.

5. Ririe KM, Rasmussen RP, Wittwer CT. 1997. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. Anal Biochem 245:154–160.

6. Gibbs RA, Nguyen PN, McBride LJ, Koepf SM, Caskey CT. 1989. Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of in vitro amplified cDNA. Proc Natl Acad Sci USA 86:1919–1923.

7. Leren TP, Rodningen OK, Rosby O, Solberg K, Berg K. 1993. Screening for point mutations by semi-automated DNA sequencing using Sequenase and magnetic beads. Biotechniques 14:618–623.

8. Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729.

9. Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186–194.

10. Kwok PY, Carlson C, Yager TD, Ankener W, Nickerson DA. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. Genomics 23:138–144.

11. Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25:2745–2751.

12. Crowe ML. 2005. SeqDoC: rapid SNP and mutation detection by direct comparison of DNA sequence chromatograms. BMC Bioinformatics 6:133.

13. Manaster C, Zheng W, Teuber M, Wächter S, Döring F, Schreiber S, Hampe J. 2005. InSNP: a tool for automated detection and visualization of SNPs and InDels. Hum Mutat 26:11–19.

14. Ngamphiw C, Kulawonganunchai S, Assawamakin A, Jenwitheesuk E, Tongsima S. 2008. VarDetect: a nucleotide sequence variation exploratory tool. BMC Bioinformatics 9: (Suppl. 12) S9.

15. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P. 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15:436–442.

16. Ellard S, Shields B, Tysoe C, Treacy R, Yau S, Mattocks C, Wallace A. 2009. Semi-automated unidirectional sequence analysis for mutation detection in a clinical diagnostic setting. Genet Test Mol Biomarkers 13:381–386.

17. Bonfield JK, Rada C, Staden R. 1998. Automated detection of point mutations using fluorescent sequence trace subtraction. Nucleic Acids Res 26:3404–3409.

18. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38:375–381.

19. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, Rowe W, Liu PP, Gibbs RA, Buetow KH. 2005. SNPdetector: a software tool for sensitive and accurate SNP detection. PLoS Comput Biol. 2005 5:e53.

FIGURE LEGENDS

Figure 1

The initial interface window, showing *GeneScreen*'s various tabs for loading reference files and test sequences, and performing analysis for the presence of mutations.

Figure 2

The *GeneScreen* results grid. Each column of the grid represents a position where one or more test sequences has a variant (red cell). The sequence around the putative variant is displayed by simply hovering the mouse over the cell. The grey cells represent a region that could not be aligned to the reference sequence.

Figure 3

Example of a web-page output format generated by *GeneScreen*. The blue and green sequence rows in the image panels show how the program deconvolutes heterozygous deletions and deduces the two allelic sequences. Note that *GeneScreen* flips the reverse strand sequence to match the orientation of the reference; consequently the superimposed staggered deletion alleles appear at the left of the image on reverse strand reads. The program includes a facility for editing and updating these web-pages to incorporate newer results.

Figure 4

Examples of migration artefacts and other sequencing anomalies that can interfere with mutation calling. See text for details.

Figure 5

Quality values of sample sequences, classified according to their number of identified sequence variants.

Licence for Publication

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in Journal of Medical Genetics and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence (http://group.bmj.com/products/journals/instructions-for-authors/licence-forms).

Figure 1

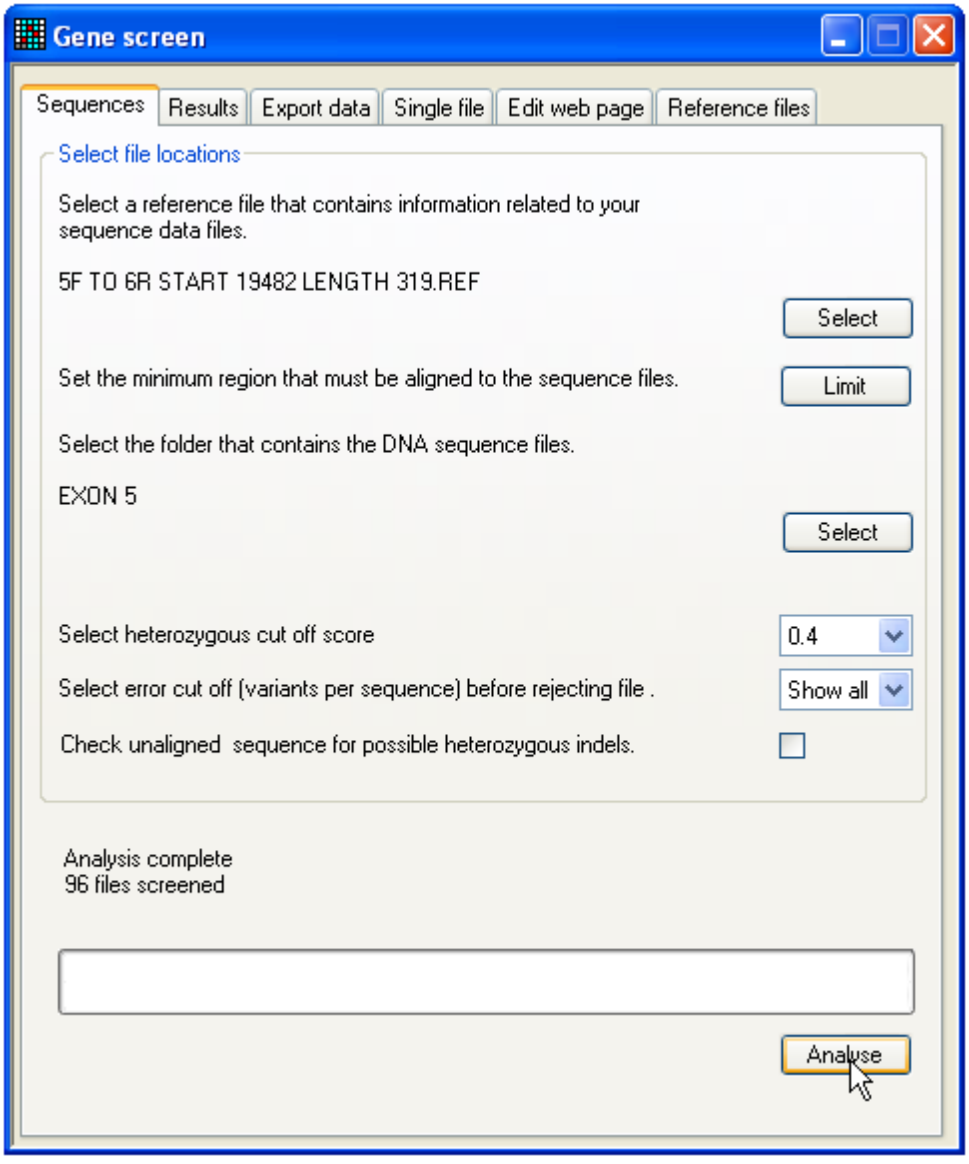

Gene screen
Sequences
Results
Export data
Single file
Edit web page
Reference files
Select file locations
Select a reference file that contains information related to your sequence data files.
5F TO 6R START 19482 LENGTH 319.REF
Select
Set the minimum region that must be aligned to the sequence files.
Limit
Select the folder that contains the DNA sequence files.
EXON 5
Select
Select heterozygous cut off score
0.4
Select error cut off (variants per sequence) before rejecting file .
Show all
Check unaligned sequence for possible heterozygous indels.
Analysis complete
96 files screened
Analyse

Figure 2

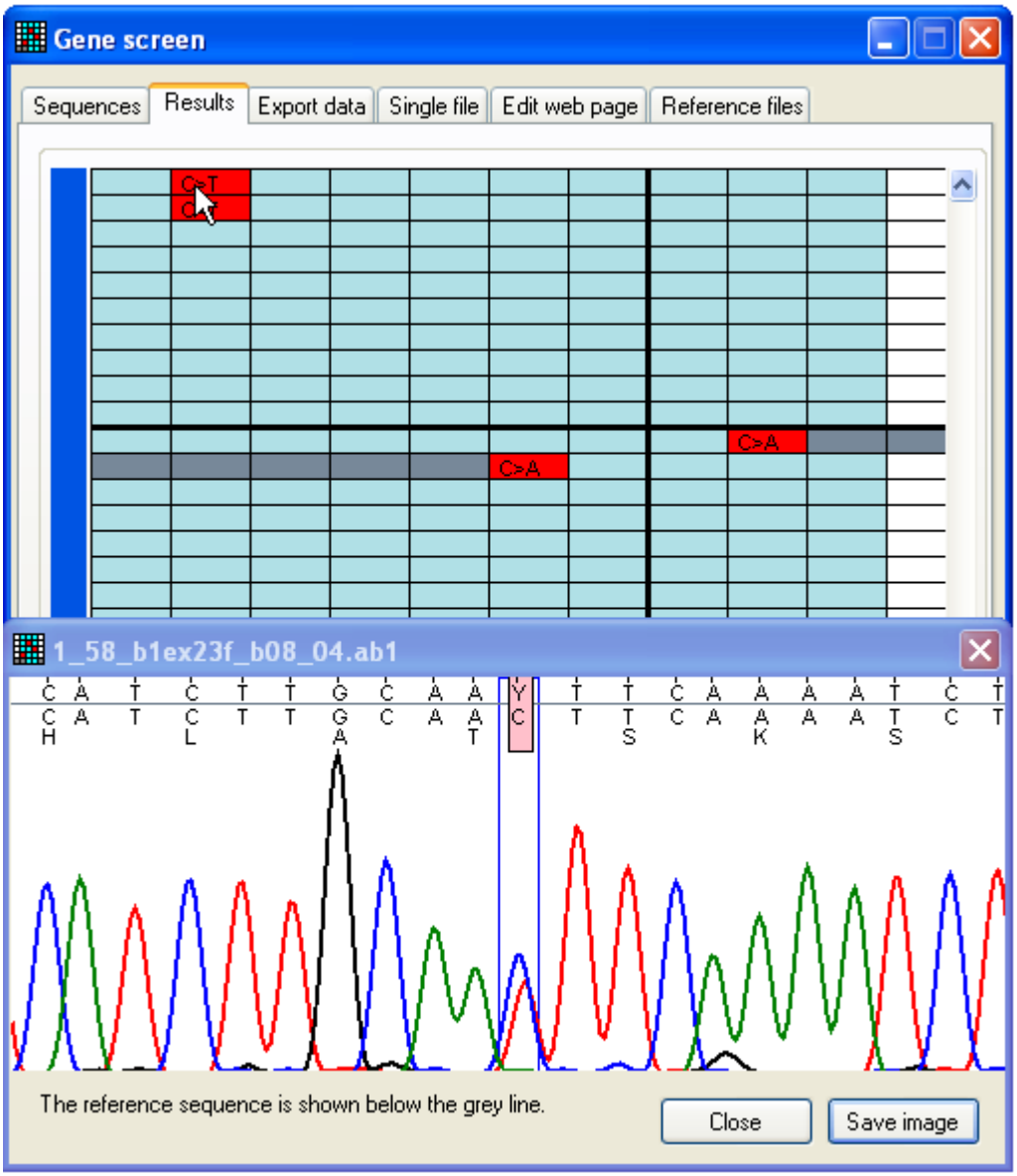

Gene screen
Sequences
Results
Export data
Single file
Edit web page
Reference files
C>T
C>A
C>A
1_58_b1ex23f_b08_04.ab1
The reference sequence is shown below the grey line.
Close
Save image

Figure 3

Allelic variations in HPGD

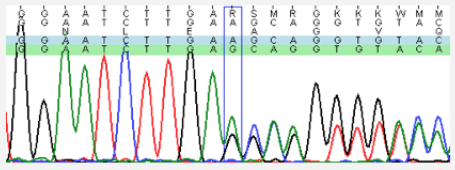
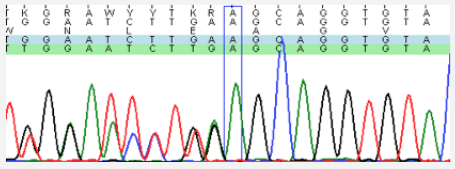
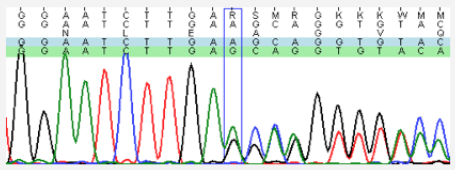
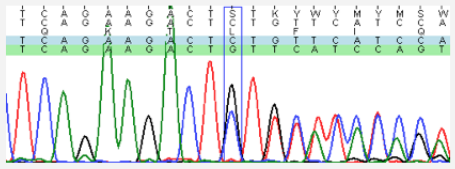
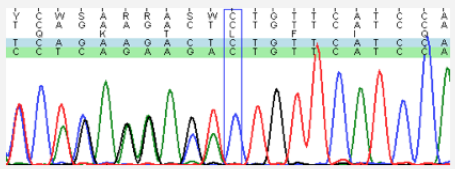
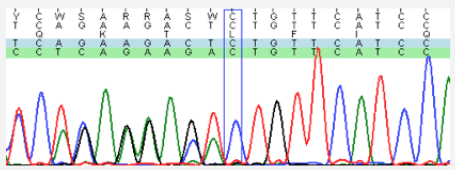
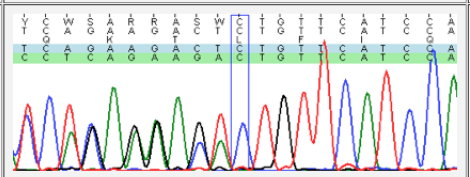
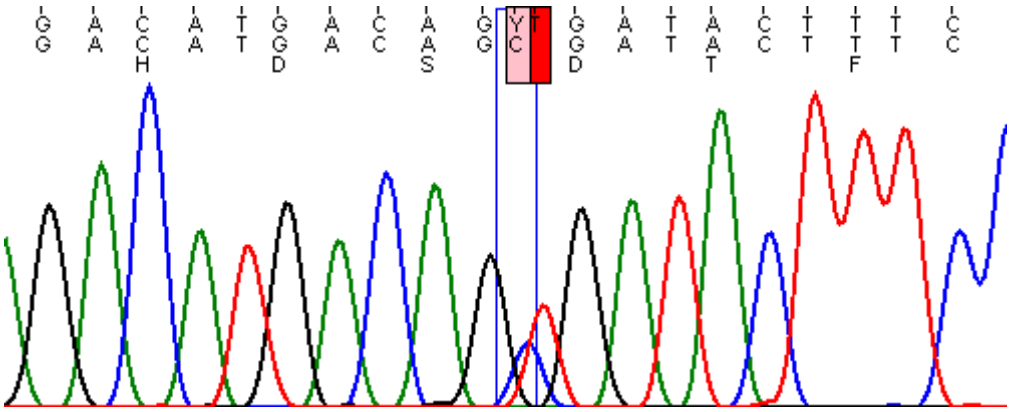
Name Ref file OMIM ID	Position	Variant	Patient ID	Image	Disease/ phenotype	Remarks
HPGD hpgd.ref 259100	562 Exon 2	g.120delA c.120delA FrameShift	[t2]_f08_2009-04-07.ab1		PHO Affected child 1	Forward strand
HPGD hpgd.ref 259100	562 Exon 2	g.120delA c.120delA FrameShift	[u2]_c09_2009-04-04.ab1		PHO Carrier father	Reverse strand
HPGD hpgd.ref 259100	562 Exon 2	g.120delA c.120delA FrameShift	[v2]_g08_2009-04-07.ab1		PHO Affected child 2	Forward strand
HPGD hpgd.ref 259100	617 Exon 2	g.175_176delCT c.175_176delCT FrameShift	[s2]_e08_2009-04-07.ab1		PHO Carrier mother	Forward strand
HPGD hpgd.ref 259100	617 Exon 2	g.175_176delCT c.175_176delCT FrameShift	[s2]_a09_2009-04-04.ab1		PHO Carrier mother	Reverse strand
HPGD hpgd.ref 259100	617 Exon 2	g.175_176delCT c.175_176delCT FrameShift	[t2]_b09_2009-04-04.ab1		PHO Affected child 1	Reverse strand
HPGD hpgd.ref 259100	617 Exon 2	g.175_176delCT c.175_176delCT FrameShift	[v2]_d09_2009-04-04.ab1		PHO Affected child 2	Reverse strand

Figure 4

A

FORWARD



B

REVERSE

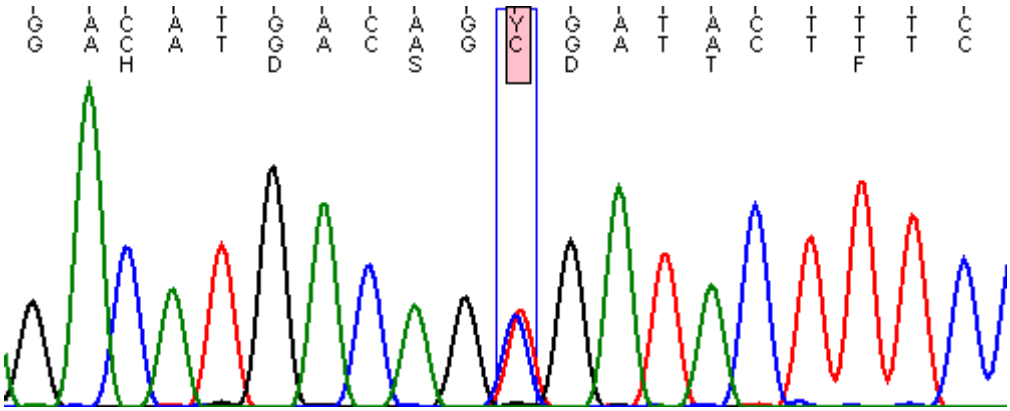
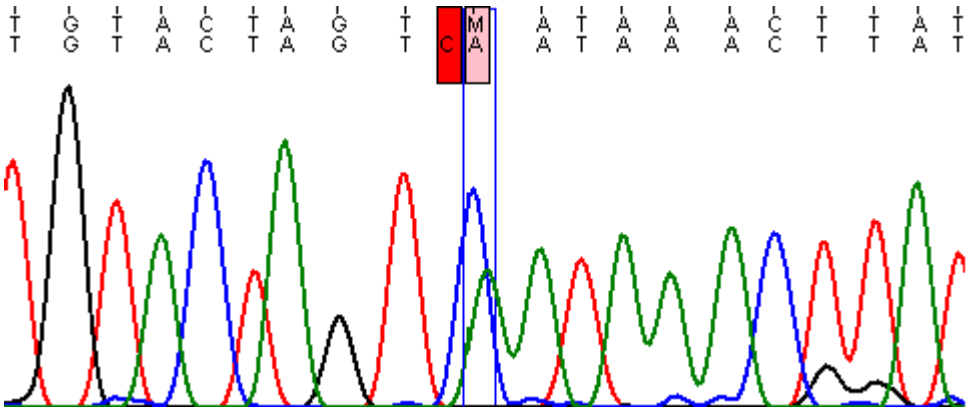


Figure 4 (cont.)

C

FORWARD



D

REVERSE

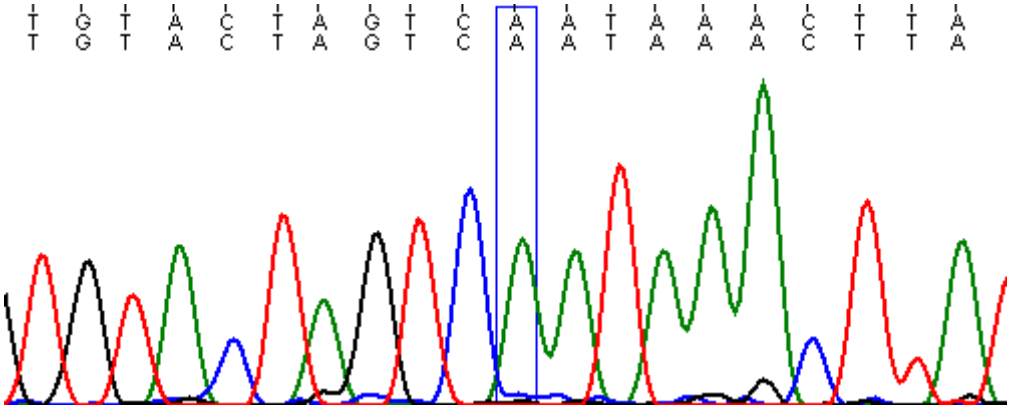
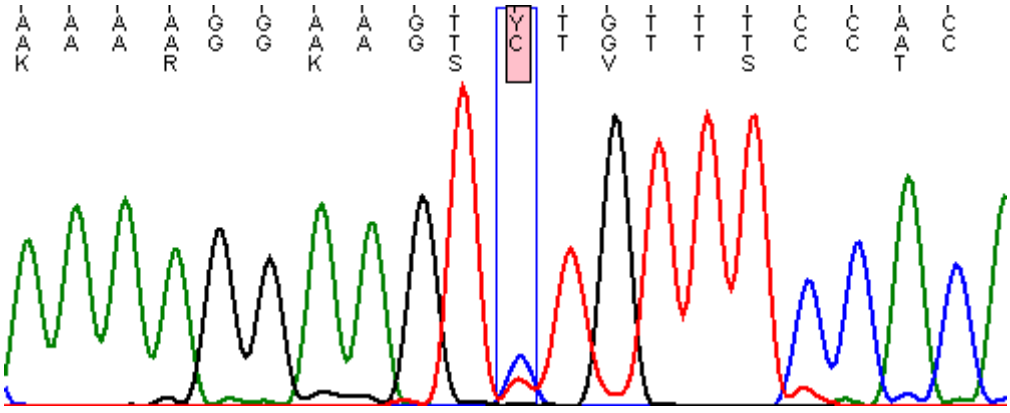


Figure 4 (cont.)

E

FORWARD



F

REVERSE

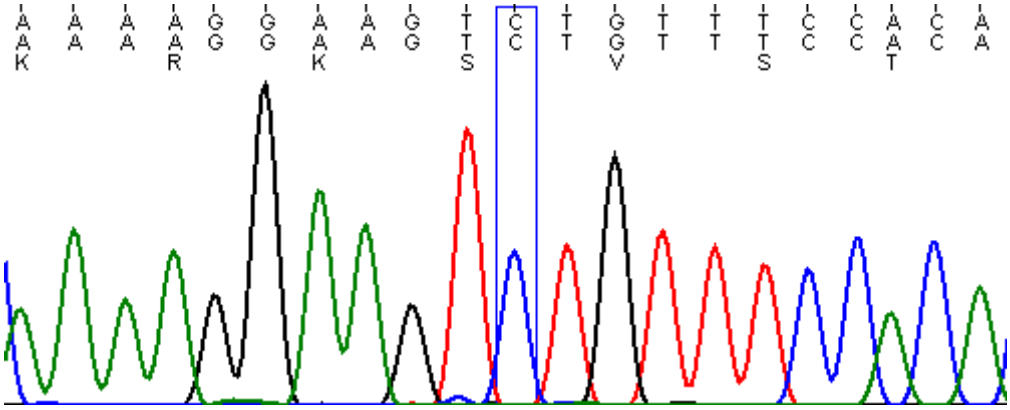
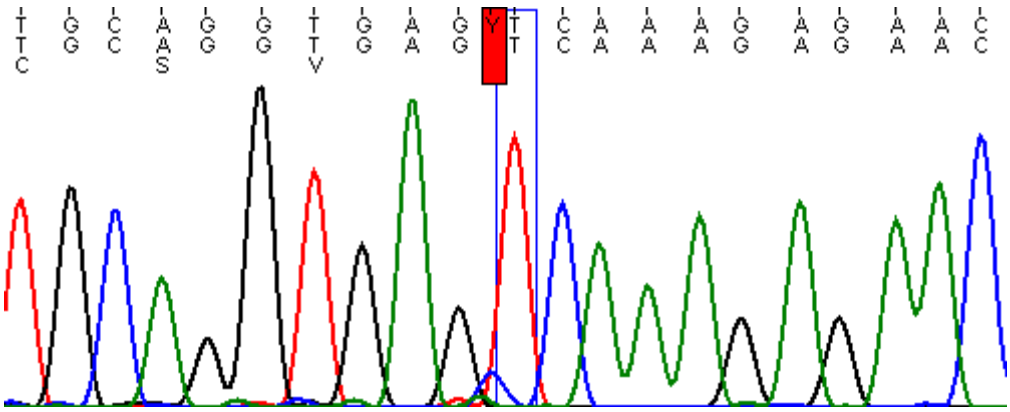


Figure 4 (cont.)

G

FORWARD



H

REVERSE

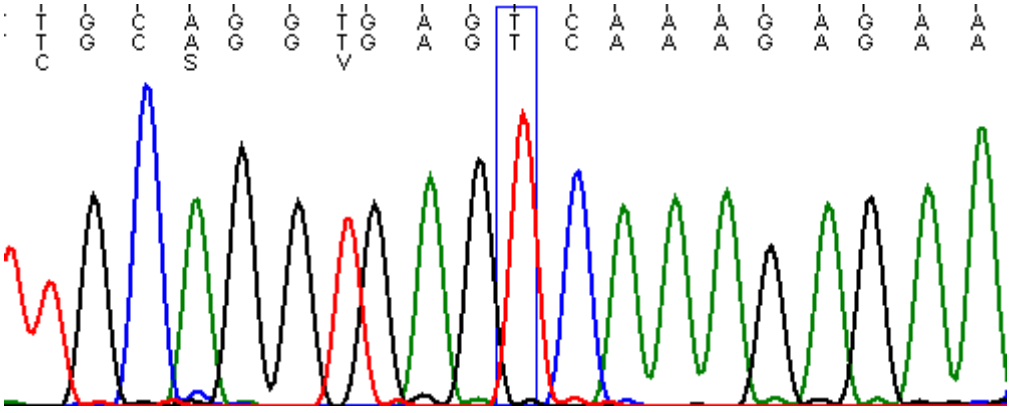
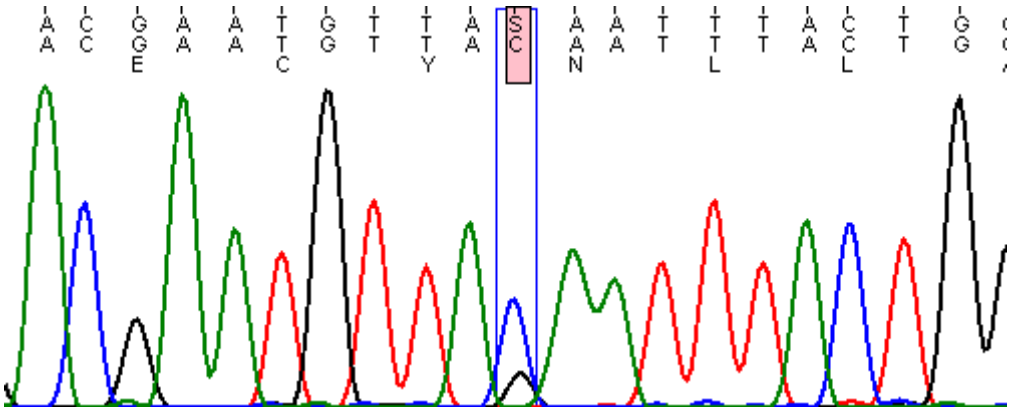


Figure 4 (cont.)

I

FORWARD



J

REVERSE

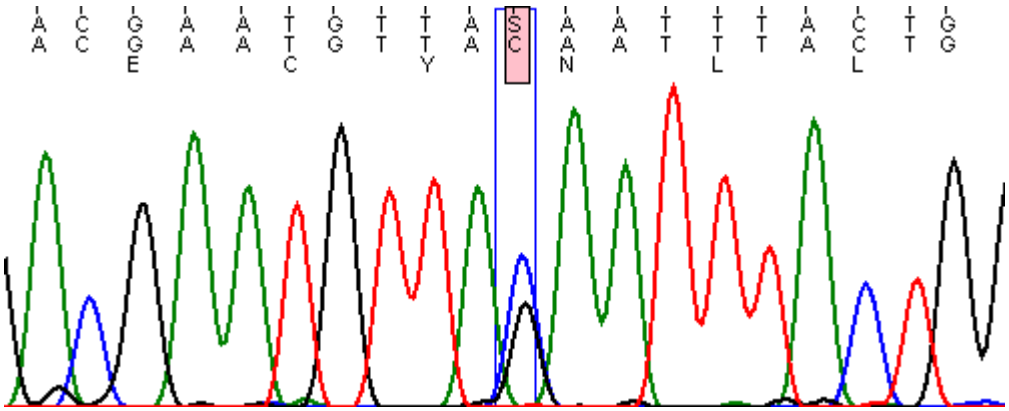
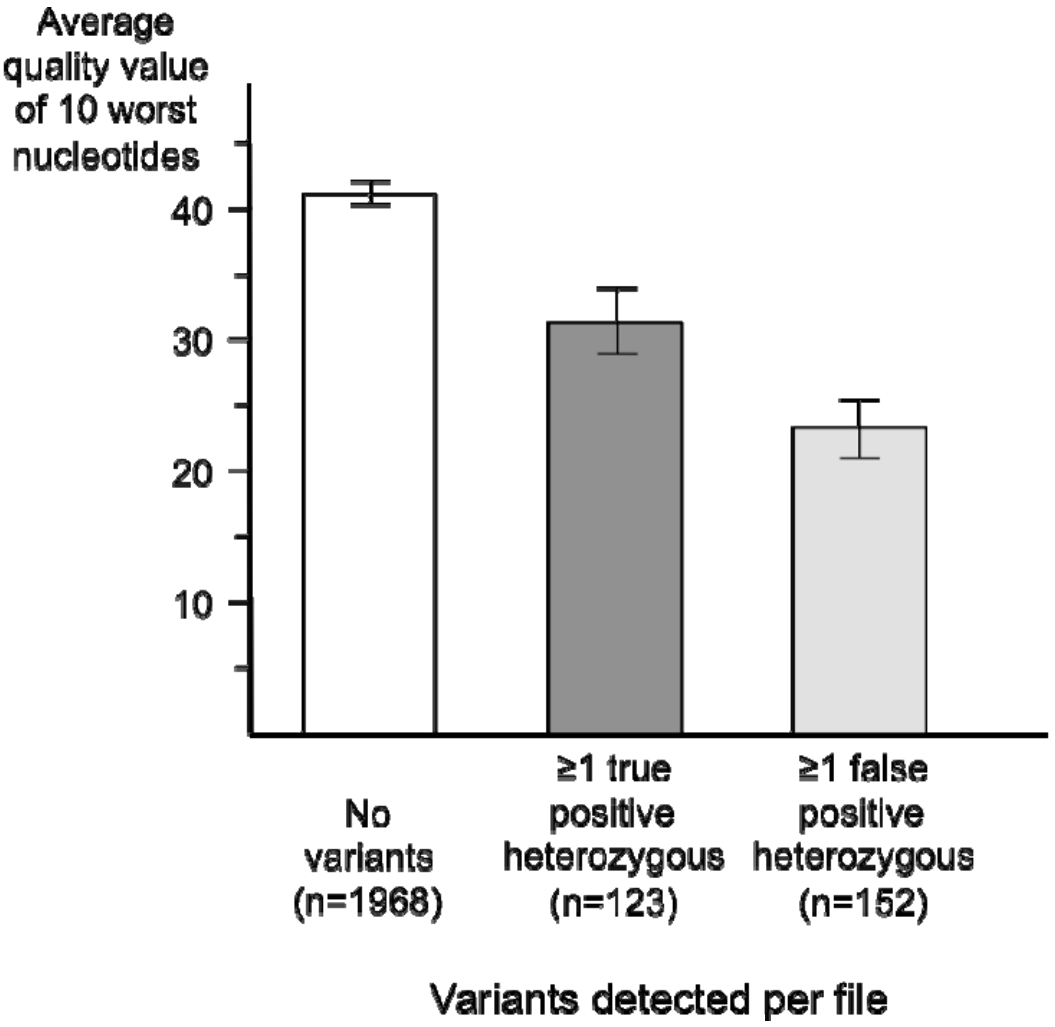


Figure 5


Average quality value of 10 worst nucleotides
40
30
20
10
No variants (n=1968)
≥1 true positive heterozygous (n=123)
≥1 false positive heterozygous (n=152)
Variants detected per file