
CHAPTER 10

SNP Discovery and PCR-based Assay Design: From *In Silico* Data to the Laboratory Experiment

ELLEN VIEUX¹, GABOR MARTH² AND PUI KWOK³

¹*Washington University School of Medicine, St. Louis, MO, USA*

²*National Center for Biotechnology Information, Bethesda, MD, USA*

³*Cardiovascular Research Institute, University of California, San Francisco, USA*

- 10.1 Introduction
- 10.2 SNP identification
 - 10.2.1 PolyBayes
 - 10.2.2 PolyPhred
 - 10.2.3 Sequencher
 - 10.2.4 Non-sequencing methods
- 10.3 PCR primer design
 - 10.3.1 Tools
 - 10.3.2 Custom primer design services
 - 10.3.3 Public databases
- 10.4 Broader PCR assay design issues
 - 10.4.1 Obtaining sequence
 - 10.4.2 Repeat masking
 - 10.4.3 Setting experimental and design parameters
- 10.5 Primer selection
 - 10.5.1 Design specific to pooled sequencing
 - 10.5.2 Design specific to Single Base Extension (SBE) reactions
- 10.6 Problems related to SNP assay validation
- 10.7 Conclusion
- References

10.1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most abundant form of DNA sequence variation in the human genome. It is widely believed that a significant fraction of SNPs contribute to our susceptibility to various diseases. In order to identify the SNPs associated with diseases, however, many groups are pursuing a case–control mapping strategy that requires a large number of SNP markers distributed throughout the human genome. Once a set of genes is implicated in a disease (either by genetic mapping or by obtaining biological evidence), the candidate genes are scanned for sequence variations that are likely to alter the genes' function. Therefore, identifying single base-pair changes, in a global or targeted fashion, is extremely important in genome research.

The central public polymorphism database, dbSNP (Sherry *et al.*, 2002), serves as an archival repository of nucleotide sequence variations. An important subset of these data, nearly 100,000 SNPs in transcribed regions, were found by analysing clusters of expressed sequence tags (ESTs) (Buetow *et al.*, 1999, 2001; Irizarry *et al.*, 2000) or by aligning ESTs to the human reference sequence (Marth *et al.*, 1999). The vast majority of genomic SNPs (single base pair variations found by analysing genomic sequence clones without regard to whether they represent exonic DNA) were discovered in sequences from restricted genome representation libraries (Altshuler *et al.*, 2000), random shotgun reads aligned to genome sequence (Sachidanandam *et al.*, 2001), and in the overlapping sections of the large-insert clones (mainly bacterial artificial chromosome, or BAC) that make up the public human reference genome (Tallion-Miller *et al.*, 1998). Because most sequences of these comparisons involved a small number of chromosomes (typically two), this collection of SNPs is enriched for common variants. Experimental characterization of these polymorphisms demonstrates that many of them occur at a high frequency in independently chosen samples, and often segregate in all or most human populations (Marth *et al.*, 2001). By the same argument, many rare polymorphisms, including those that cause noticeable but rare phenotypic effects, are likely to be absent from this set. The identification of rare phenotypic mutations will require significantly higher sample sizes and may only be possible by the cross-comparison between large samples of affected patients and those of controls (see Halushka *et al.* (1999) for an example of such a study).

Because the numbers are extremely large and the need for identifying SNPs in a timely fashion is great, computer tools are indispensable in the SNP discovery process. Fundamentally, one identifies a SNP by comparing two or more sequences from the same region on the chromosome. This can be done quite easily if the DNA sequence quality is high and the sequence data are derived from cloned DNA because each clone comes from a single copy of one of the two chromosomes in the diploid human cell. There are no unambiguous bases in regions where the data quality is high. In the case of identifying SNPs in targeted regions in the genome, one amplifies genomic DNA by PCR and sequences the PCR products derived from different individuals. In this situation, SNP discovery is complicated by the fact that the same regions on both chromosomes in the diploid cell are amplified by PCR and some bases will be heterozygous in one or more individuals. A good computer tool will be able to identify a SNP even when only heterozygotes and homozygotes of just one of the two alleles are present in the samples sequenced. This is not a trivial problem because the commonly used dye terminator-based DNA sequencing methods yield peaks of uneven heights at the polymorphic sites and the base-calling algorithm will frequently miscall the base at these sites in the sequences of heterozygous individuals.

In this chapter, we will survey the computer tools used in global and targeted SNP discovery and PCR-based assay design. Instead of describing the mechanics of how to use

these bioinformatic tools, we refer the reader to the primary literature and the excellent documentations of these tools and concentrate on explaining the approaches these tools take and the limitations (if any) they may have.

10.2 SNP IDENTIFICATION

Computational discovery of polymorphisms in sequence data usually follows a four-step procedure. First, sequences of high similarity in multiple individuals are identified, usually with a BLAST (Altschul *et al.*, 1990) similarity search. To avoid spurious similarity due to known human repeats, sequences are masked for high copy number repetitive elements with REPEATMASKER (Arian Smit, unpublished data). Still, the possibility exists that the sequences originate from regions of as yet uncharacterized chromosomal duplications (Lander *et al.*, 2001). Inclusion of a second, paralogue-filtering step into the procedure can reduce false positive SNP predictions arising from comparing paralogous sequence copies. Following this step, false predictions due to paralogy were as low as 0.2% of the data collected through pooled SNP characterization in the Kwok laboratory (unpublished data).

The third step is the construction of a base-wise multiple alignment of the sequences. In the general case, this is a computationally expensive task. Aligning expressed sequences is even more complicated because of exon–intron punctuation and possible alternative splice variants. In the case of human data one can organize fragmentary sequences on top of the nearly complete reference sequence (Lander *et al.*, 2001). This approach was shown to work well for discovering SNPs in clusters of cDNA sequences (Marth *et al.*, 1999).

In the fourth (and final) step, sequences in the precise, base-to-base multiple alignment are scanned for nucleotide differences. Because of the possibility of sequencing errors, not every mismatch is a polymorphic site. Discrimination between true polymorphism and sequencing error uses statistical tools based on measures of sequence accuracy, or base quality values (Ewing and Green, 1998; Ewing *et al.*, 1998). Each SNP prediction is accompanied by a measure of confidence. Accurate confidence values permit one to use the highest number of candidates with an acceptable false positive rate.

Both commercial and academically developed programs are available for use in SNP detection. Some methods use sequence quality data to eliminate false positives due to poor sequencing quality. Others incorporate expected mutation rates to distinguish true SNPs. The most prominent methods of detecting SNPs are PolyBayes, PolyPhred, and Sequencher. Other methods incorporate neighbourhood quality standard (NQS) generated by Phred (Ewing *et al.*, 1998) to determine the quality of the data surrounding the SNP (Altshuler *et al.*, 2000; Mullikin *et al.*, 2000). PolyPhred and PolyBayes are freely available to academic groups, while Sequencher is produced by Gene Codes Corporation (URL: www.genecodes.com). Other companies have developed software based on the same principles. Typically, these products either offer a built-in graphical interface, or use an external, licensable interface program (such as CONSED). They can be used for both comparison of short known regions, or long shotgun regions, and are extremely useful when searching known regions of interest for novel SNPs.

10.2.1 PolyBayes

The POLYBAYES program was developed for *de novo* SNP discovery in non-ambiguous (clonal) sequence data (Marth *et al.*, 1999). The SNP detection algorithm employs a Bayesian approach to combine prior knowledge (such as average polymorphism rate or expected transition to transversion ratio) with the base calls and base quality values of the sequences in the multiple sequence alignment. Each SNP prediction comes with a

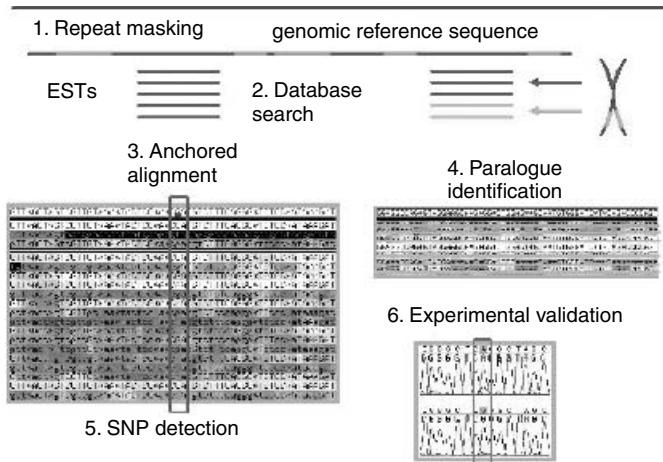


Figure 10.1 The POLYBAYES SNP discovery tool applied to EST-Mining. (1) The genome sequence (BAC clone or assembled sequence contig) is masked for known, large copy number human repeats. (2) Using the BLAST similarity search tool, expressed sequences in the public database (dbEST) that match the genomic sequence are identified. (3) Matching ESTs are aligned to the genomic sequence using an anchored alignment approach. (4) Possible paralogous ESTs are identified and discarded. (5) The multiple alignment is scanned for polymorphic sites. (6) Candidates are validated by sequencing in independent, population-specific DNA pools.

predicted true positive rate, or ‘SNP score’, which have been shown to be accurate (Marth *et al.*, 2001). Figure 10.1 illustrates an example of using POLYBAYES for SNP discovery in ESTs aligned to genome sequence. POLYBAYES has been used to discover SNPs in overlapping regions of human BACs (Marth *et al.*, 1999), in *C. elegans* (Wicks *et al.*, 2001) and *Drosophila* (Berger *et al.*, 2001).

10.2.2 PolyPhred

PolyPhred was developed to be used with Phred, Phrap, and CONSED to identify candidate SNPs in sequence trace data (Ewing and Green, 1998; Gordon *et al.*, 1998; Table 10.1). In Consed, coloured marks in the sequence alignment are used to indicate candidate SNPs as well as confidence in the variations base call. The accuracy of the calls by PolyPhred has been tested using previously screened mitochondrial DNA. The results show that this software exhibits over 95% accuracy depending on the quality of the sequence traces (Nickerson *et al.*, 1998). The primary use of PolyPhred is to identify SNPs in PCR-amplified data as it can detect heterozygous sequence peaks, and is thus widely employed in sequence-based genotyping applications.

10.2.3 Sequencher

Sequencher is a tool developed by GeneCodes for sequence alignment, annotation, editing and mutation identification. Although it is a commercial product, a free demo version is available (www.genecodes.com/features/html). Sequencher can be used with automated sequencers such as ABI, Pharmacia/ALF, LI-COR and VISTRA. GeneCodes continues to

TABLE 10.1 Tools and Related Resources for Primer Design**SNP detection tools**

Sequencher	http://www.genecodes.com/features.htm
PolyPhred	http://droog.mbt.washington.edu/PolyPhred.html
POLYBAYES	http://www.genome.wustl.edu/gsc/polybayes/

Repeat masking tools

RepeatMasker	www.genome.washington.edu/uwgc/analysistools/repeatmask.html
MaskerAid	http://sapiens.wustl.edu/maskeraid/

Primer design tools

Primer3	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
TSC primer db	ftp://snp.cshl.org/pub/SNP/
Primer design tips	http://www.alkami.com/primers/refdsgn.htm

Tools for sequence extraction and manipulation

SNPper	http://bio.chip.org:8080/bio/
UCSC HGB	http://genome.ucsc.edu/index.html

add new tools and functionality to the software. The most appealing part of the software is the graphical interface, which is intuitive and easy to use on multiple platforms.

10.2.4 Non-sequencing Methods

Several groups have explored non-sequencing methods for SNP discovery. Among the most promising of these techniques is the use of high density DNA chips (Dong *et al.*, 2001). Variations of this method have been used to scan for genome wide SNPs (Wang *et al.*, 1998), in mitochondrial DNA (Chee *et al.*, 1996), and to scan all of chromosome 21 (Patil *et al.*, 2001). Methods for SNP scanning using DNA chips vary considerably in design (for a review see Draghici *et al.*, 2001).

10.3 PCR PRIMER DESIGN

A large number of candidate SNPs exists in public databases. Key to taking advantage of this resource is the ability to design PCR assays to amplify these loci uniquely and SNP genotyping assays for genetic studies under standardized conditions. Genetic researchers wanting to validate and assay SNPs are faced with the need for high throughput primer design. Manual picking of primers is time consuming, and some automated tools only allow for submission of one sequence at a time. There are many tools available over the web as well as software. In addition, some commercial companies offer genotyping assay design by order for their customers and a few assays are available through public databases.

10.3.1 Tools

Currently there is no standard method for calculating the annealing temperature (TM) of primers. Although many tools have been developed to determine the annealing temperature, their results vary. Furthermore, many of these programs use different entropy and enthalpy tables in their TM calculations, leading to further discrepancies (Owczarzy *et al.*, 1997). Despite these variances most of these tools will work and one program that

has become a standard is Primer3 (http://www-genome.wi.mit.edu/genome_software/other/primer3.html). A comprehensive review of primer picking and TM predicting tools can be found at <http://www.alkami.com/primers/refdsgn.htm>.

Primer3 is a standard because it is freely available and easy to use. It is particularly useful for high throughput design because it can determine primers for multiple sequences at once. Some of its particular strengths are its many useful and well-documented options, its easily parsed output and its simple command line interface. Primer3 can be used for the design of both PCR primers and internal sequencing primers. Although Primer3 allows for individual SNP position targets and target lengths to be set for each sequence, if the data is highly varied in position and length it is possible to avoid setting parameters for each SNP by pre-formatting the data. The SNP sequence retrieval option on SNPper (see below) is a tool that can provide this uniformly formatted flanking sequence.

10.3.2 Custom Primer Design Services

Although primer design may be carried out in-house, many companies as well as public databases are offering high throughput design as part of their product support. Sequenom is one such example. Sequenom is in the process of making primers available through a site called RealSNP (www.RealSNP.com). Applied Biosystems is another company providing primer design through their 'Assay by DesignTM' Genomic Assay Service. The researcher provides the sequence, while Applied Biosystems designs and test all assays. These designs are optimized for TaqmanTM assays (<http://www.appliedbiosystems.com/>). There is no charge if an assay cannot be designed. Perkin-Elmer will also be providing SNP-specific assays through their website.

10.3.3 Public Databases

Primers generated by The SNP Consortium (TSC) Allele Frequency Project are available via ftp (Table 10.1). These primers have been released, by some of the groups, to the public with the assistance of TSC. It should be noted that these primers have been generated by separate groups via different methods and for specific experimental conditions. The NCBI's dbSNP database also contains primer designs for some of the SNP entries, but these have not been specifically designed for SNP validation (Sherry *et al.*, 2002). In addition to these public databases the Kwok laboratory has over 980,000 assays designed for sequencing of PCR products, for the specific purpose of pooled allele frequency determination. The Kwok laboratory also maintains over 1,400,000 SNP genotyping assays designed for single base extension using fluorescence polarization detection (available at <http://snp.wustl.edu>).

10.4 BROADER PCR ASSAY DESIGN ISSUES

SNP assay methods have three major components: (1) allelic discrimination methods, (2) reaction formats and (3) detection methods. Each area presents different challenges during SNP assay design. The most important consideration for assay design is the method of allelic discrimination. These methods vary greatly. For example four main methods of allelic discrimination are allele-specific hybridization, primer extension (includes single base extension), ligation and invasive cleavage. The reaction formats are either homogeneous reactions or solid phase reactions and the detection methods currently use product light emissions, product-mass measurements and electrical property changes in the product (Kwok, 2001).

In some cases the critical parameters that apply to one technique will not apply to others. However, almost all SNP genotyping assay techniques use PCR to amplify DNA. In order to design the correct primers one must first determine the method of assay. However, there are some basic guidelines used when designing primers for genomic sequence. All designs require obtaining sequence, repeat masking, setting experimental and design parameters, picking primers and formatting the information.

10.4.1 Obtaining Sequence

The flanking sequences for each SNP can be obtained from a variety of sources. For known SNPs two public databases, dbSNP and SNPper (Riva and Kohane, 2001; Table 10.1) provide a method for obtaining sequence. SNPper is run by Harvard's Children's Hospital Informatics Program (CHIP). Both dbSNP and SNPper offer batch query modes and return sequence in FASTA format. In the case of single SNP analysis, SNPper provides a link to the Primer3 website which will import the retrieved sequence into Primer3 and analyse it using default values. For SNPs that can be uniquely mapped SNPper can provide up to 1000 bases on either side of the SNP. It should be noted that at this time SNPper does not contain the most recent uniquely mapped SNPs in dbSNP.

10.4.2 Repeat Masking

A large amount of the genome consists of repeated regions or low complexity DNA. It is important to avoid selecting primers from these regions in order to avoid amplification of multiple products. Masking the repeats or making repeated sequence unavailable to the automated primer-picking programs prevents most unwanted amplification. A commonly used program for masking is RepeatMasker (see Table 10.1). A new resource that improves upon RepeatMasker is MaskerAid (Table 10.1), which increases the speed of masking more than 30-fold (Bedell *et al.*, 2000). Default parameters in RepeatMasker will mask known repeat regions with Ns. RepeatMasker accepts FASTA files, and returns the sequence in the same format. Ready masked sequence can also be obtained from some of the public databases. dbSNP provides sequence in FASTA format with low-complexity sequence in lower case, while the University of California Santa Cruz Human Genome Browser (UCSC HGB) has options to save repeats as either Ns or lower case. However, this format is problematic when trying to represent the start and stop exons and introns on UCSC HGB, because lower case can also be used to represent introns. In some cases two files may be required to represent the masked and unmasked forms of sequence.

Masking repeats can only be accomplished in known repeat regions with current resources. However, there remain repeat regions of the genome that have not yet been identified. By using pooled sequencing, it is possible to identify regions that have duplicated and subsequently diverged. These can be identified by the presence of a large number of apparent SNPs that are all 50% in frequency, as shown in Figure 10.2. When designing SNP specific primers within PCR products, for example for a single base extension assay, the RepeatMasking stage is not necessary.

10.4.3 Setting Experimental and Design Parameters

If a large number of SNP candidates are to be assayed it is more efficient to eliminate the experiments that are less likely to be successful *in vitro* during the *in silico* design stage. Stringent design parameters allow for a first level of screening when designing primers.

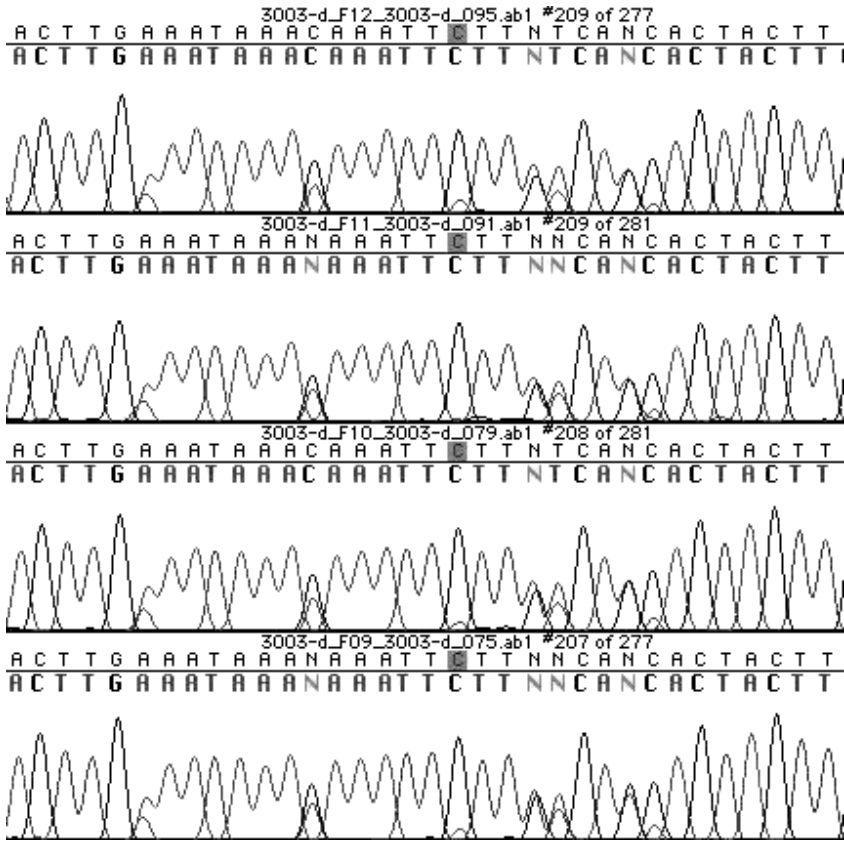


Figure 10.2 Duplicated and diverged regions in genome characterized by multiple ‘SNPs’ at 50% frequency in each pool.

Primer design programs such as Primer3 allow for input of both experimental parameters and primer structure parameters. The second level of screening can be done after candidate primers are chosen by primer selection programs to determine if the primers are likely to work.

Some suggestions for optimizing design parameters for the best experimental results can be found in *PCR Applications* (Beasley *et al.*, 1999). In general most primer design methods work (see http://gsu.med.ohio-state.edu/primer_design/sld001.htm for a detailed guide). However, under stringent experimental conditions optimized design parameters can decrease the level of experimental failures.

10.5 PRIMER SELECTION

In most design programs primer picking is preformed by the software. The primers are picked according to the specified parameters. If more than one primer set is returned some post processing will be required to select the most appropriate pair. Post processing

can also be necessary for techniques such as pooled sequencing, where selection of a sequencing primer from the PCR primers is required.

10.5.1 Design Specific to Pooled Sequencing

Pooled sequencing uses sequencing to observe the frequency of a SNP in a group of individuals in one reaction. The candidate SNP and its flanking sequence is amplified from pools of DNA each containing individuals and a single reference individual. After sequencing of the PCR products is performed using fluorescent dye-terminators, the sequence traces are aligned, allowing the allele frequencies to be estimated (Kwok *et al.*, 1994). Pooling DNA in this way prior to PCR amplification and estimating allele frequencies by subsequent quantitation of trace peak heights yields considerable time and cost savings.

There are several steps to designing pooled sequencing reactions. This method of design is carried out on a UNIX-based system, using RepeatMasker and Primer3. Repeats are masked before choosing PCR primers. Sequence that is not masked is retained for post processing. The input for Primer3 is set according to the optimized parameters (Beasley *et al.*, 1999) with a few optimizations. The optimizations are most important in the placement of the primers relative to the SNP. The primers are not allowed closer than 25 bases to the SNP, but are close enough to use one of the PCR primers for sequencing. After running Primer3 the results are processed to select for the best sequencing primer based on criteria to optimize experimental performance. These criteria are (1) the sequencing primer should be 100 bases from the target and (2) there should be no poly As or Ts greater than eight bases and no poly GTs or CAs greater than 10 pairs between the primer and the SNP. This design has been shown to work with less than 3% experimental failure and allows for the primers to be far enough from the SNP that the sequence is of high quality around the target as shown in Figure 10.3. During the design process as many as half of the SNPs fail to meet the design criteria, but this failure is at far lower cost than laboratory-based trial and error (Vieux *et al.*, 2002).

10.5.2 Design Specific to Single Base Extension (SBE) Reactions

SBE requires a primer that abuts the SNP under test. The primer is then extended by a single base, usually a labelled ddNTP (Hsu *et al.*, 2001). By using two different labels for the ddNTPs representing the two possible alleles, the allelic state of the SNP can be determined. SNP-specific SBE primer design can be undertaken using many of the same tools as pooled sequencing primer design. Both require repeats to be masked before designing PCR primers. The SNP-specific primers are chosen using non-masked sequence. The PCR product sizes can be smaller than for sequencing for all of the single base extension reactions. The SBE primer should not hang over the end of the PCR product and the PCR primers should not overlap with the SBE primer. In Primer3 the primers can overlap the target so it is important to give a SNP a large enough target area to prevent the overlap of primers. When choosing parameters and methods for SBE primers it is important to remember that different methods can have different primer requirements.

We have found that picking the shortest primer from 16–40 bases which has a TM between 60–65 degrees works well. In order to calculate TM for a small number of SBE primers it is possible to use free tools on the web. For high throughput design the best option is to solve TM equations after determining which set of entropy and enthalpy tables work best for the relevant method (Owczarzy *et al.*, 1997), and picking the shortest primer in the defined range. Further optimization can be achieved by picking the SBE primer with the least amount of secondary structure, and fewest runs of poly As and Ts.

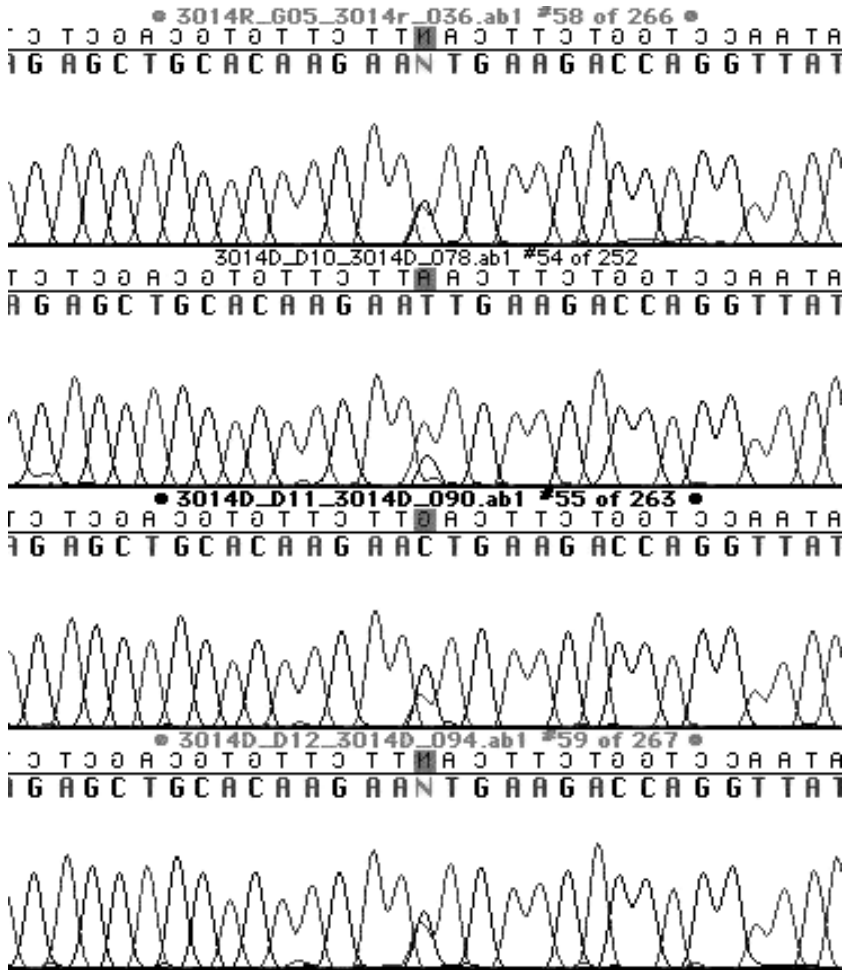


Figure 10.3 A clean pooled sequencing assay result shows a clear polymorphic site at the position of the candidate SNP.

10.6 PROBLEMS RELATED TO SNP ASSAY VALIDATION

As with any experimental design, assays for validation of candidate SNPs require attention to detail. Problems arise that are not always obvious or clearly stated in the documentation associated with the tools being used. Some problems are easy to overcome, while others cannot yet be solved.

With the completion of a final version of the human genome assembly a number of problems will be resolved, while inherent challenges will remain. There are many errors due to incorrect physical map order, gaps in physical map data and incorrect assembly (DeWan *et al.*, 2002). These errors lead to SNPs mapping to multiple locations, incorrect haplotypes and difficulty in identifying paralogues. However, SNP locations are continually amended as assemblies are progressively corrected. Map locations will continue

to change until the Human Genome Project is complete. This can cause difficulties in the analysis of data and obtaining guide sequence. Another difficulty with the unfinished map is unidentified paralogues. A SNP can appear to map to a unique position, when it is actually an artefact generated from an unknown paralogue of the original reference sequence. An example of such an artefact generated by pooled sequencing data is shown in Figure 10.2.

Guide sequence is provided for known SNPs through dbSNP, TSC and SNPper. The first two sites only provide a small amount of flanking sequence in their database for any given SNP. This can lead to failure in the design of PCR primers due to limited sequence information. SNPper provides far more flanking sequence by mapping the SNP location and retrieving guide sequence from the human genome assembly at the UCSC (Table 10.1).

Other problems are inherent when working with DNA and the current technologies. Long runs of a single nucleotide can cause sequencing reactions to fail, while insertion/deletion events can cause problems with sequencing and with SNP allelic discrimination methods such as allele-specific hybridization, primer extension (including SBE), ligation and invasive cleavage. These problems may only be solved with new technologies for SNP characterization.

10.7 CONCLUSION

Given the large number of SNPs in the human genome and the potential for large-scale experimentation, bioinformatics tools are essential for SNP discovery and genotyping assay development. The tools for comparing cloned (and hence homozygous) sequences are well developed and have proven useful. However, tools for comparing genomic sequences amplified by PCR, which are often heterozygous, still have room for computational and technical improvements. Following SNP discovery, there are many assay methods for genotyping, but none can satisfy all requirements. The basic methods for assay design are well defined, but specific optimizations are different for each method. With technology improvements, some of the current problems in SNP assay design will be solved, resulting in a reduction in the number of SNPs that are refractory to successful assay design. But for now design optimization using currently available tools and careful interpretation of subsequent results will provide assays and allele frequencies for a large portion of the SNPs currently available.

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, *et al.* (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.
- Beasley EM, Myers RM, Cox DR, Lazzaroni LC. (1999). Statistical refinement of primer design parameters. In Michael DHG, Innis A, Sninsky JJ. (Eds), *PCR Applications*. Academic Press: New York, pp. 55–71.
- Bedell J, Korff I, Gish W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**: 1040–1042.

- Berger J, Suzuki T, Senti K, Stubbs J, Schaffner G, Cickson BJ. (2001). Genetic mapping with SNP markers in *Drosophila*. *Nature Genet* **29**: 475–481.
- Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, *et al.* (2001). High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci USA* **98**: 581–584.
- Buetow KH, Edmonson MN, Cassidy AB. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet* **21**: 323–325.
- Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, *et al.* (1996). Accessing genetic information with high density DNA arrays. *Science* **274**: 610–626.
- DeWan AT, Parrado AR, Matise TC, Leal SM. (2002). The map problem: a comparison of genetic and sequence-based physical maps. *Am J Hum Genet* **70**: 101–107.
- Dong S, Wang E, Hsie L, Cao Y, Chen X, Gingeras TR. (2001). Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res* **11**: 1418–1424.
- Draghici S, Kulin A, Hoff B, Shams S. (2001). Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Curr Opin Drug Discov Devel* **4**: 332–337.
- Ewing B, Hillier L, Wendl MC, Green P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175–185.
- Ewing B, Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186–194.
- Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**: 195–202.
- Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, *et al.* (1999). Patterns of screening of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet* **22**: 239–247.
- Hsu TM, Chen X, Duan S, Miller RD, Kwok PY. (2001). Universal SNP genotyping assay with fluorescence polarization detection. *Biotechniques* **31**: 560, 562, 564–568, *passim*.
- Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, *et al.* (2000). Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet* **26**: 233–236.
- Kwok P-Y. (2001). Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**: 235–258.
- Kwok P-Y, 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.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, *et al.* (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genet* **23**, 453–456.
- Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, *et al.* (2001). Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet* **27**, 371–372.
- Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, *et al.* (2000). An SNP map of human chromosome 22. *Nature* **407**: 516–520.

- Nickerson DA, Rieder MJ, Taylor SL, Tobe VO. (1998). Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res* **26**: 967–973.
- Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS. (1997). Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers* **44**: 217–239.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- Riva AA, Kohane IS. (2001). A web-based tool to retrieve human genome polymorphisms from public databases. *Proc AMIA Symp* 558–562.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- Sherry ST, Ward MH, Kholdov M, Baker J, Phan L, Smigielski EM, *et al.* (2002). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.
- Tallion-Miller P, Gu Z, Li Q, Hiller L, Kwok PY. (1998). Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res* **8**: 748–754.
- Vieux EF, Kwok P-Y, Miller RD. (2002). Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* (in press).
- Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1081.
- Wicks SR, Yeh RT, Gish WR, Waterston RH, Plasterk RHA. (2001). Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nature Genet* **28**: 160–164.