Nucleic Acids Research, 1998, Vol. 26, No. 4 967–973

Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome

Mark J. Rieder*, Scott L. Taylor, Vincent O. Tobe and Deborah A. Nickerson

Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730, USA

Received October 29, 1997; Revised and Accepted December 17, 1997

ABSTRACT

Diagnostic re-sequencing plays a central role in medical and evolutionary genetics. In this report we describe a process that applies fluorescence-based re-sequencing and an integrated set of analysis tools to automate and simplify the identification of DNA variations using the human mitochondrial genome as a model system. Two programs used in genome sequence analysis (Phred, a base-caller, and Phrap, a sequence assembler) are applied to assess the quality of each base call across the sequence. Potential DNA variants are automatically identified and 'tagged' by comparing the assembled sequence with a reference sequence. We also show that employing the Consed program to display a set of highly annotated reference sequences greatly simplifies data analysis by providing a visual database containing information on the location of the PCR primers, coding and regulatory sequences and previously known DNA variants. Among the 12 genomes sequenced 378 variants including 29 new variants were identified along with two heteroplasmic sites, automatically detected by the PolyPhred program. Overall we document the ease and speed of performing high quality and accurate fluorescence-based re-sequencing on long tracts of DNA as well as the application of new approaches to automatically find and view DNA variants among these sequences.

INTRODUCTION

The identification of DNA variations is playing an increasingly important role in probing the relationship between human genotype and phenotype. Because its complete nucleotide sequence is known (16569 bp; 1), the mitochondrial genome provides a unique system to design new approaches to identify DNA variations and to analyze the functional consequences of these in human populations.

Mitochondrial DNA (mtDNA) composes a second genome that encodes essential proteins for the oxidative phosphorylation pathway. It is inherited maternally and exhibits a rapid evolutionary rate (2). The mitochondrial genome is divided into two distinct regions: (i) the D loop (or control region), an ~1100 bp non-coding region; (ii) a functional region which encodes 13 proteins, 12S and 16S rRNA and 22 tRNAs. The analysis of sequence variation in the D loop has been applied in the study of evolutionary genetics (3, 4) and in forensic situations (5–7), while identification of single nucleotide mutations in the mitochondrial coding region has been associated with various maternally inherited or sporadically occurring pathologies (8,9).

Advances in the polymerase chain reaction (PCR) and DNA sequencing methods have made it feasible to study the mtDNA sequence of numerous individuals (10–12). Regional and whole genome amplification of mtDNA (13) using PCR has greatly simplified the process of comparing sequences to identify nucleotide variations and has eliminated the need for cloning. Many approaches have been used to comparatively scan mtDNA for sequence variants, including denaturing gradient gel electrophoresis (14), chemical or enzymatic treatment (15), single-stranded DNA conformational analysis (16) and hybridization with allele-specific oligonucleotides (17) or on oligonucleotide arrays (18), as well as direct DNA sequencing of specific nucleotides (19) or regions (6). Among these methods direct sequence analysis has many advantages because it provides complete information about the location and nature of any DNA variants in a single pass, is amenable to automation, widely available and simple to apply (only a single set of reagents and assay conditions are required).

In this report we present an approach for automating the detection of sequence variants in human mtDNA which can be easily adapted to identify and analyze DNA polymorphisms in long tracts of nuclear DNA of biological or medical interest. This approach integrates the use of conventional fluorescence-based sequencing with DNA analysis software that measures sequence quality and improves the accuracy and automation of detecting DNA variations.

MATERIALS AND METHODS

PCR primers

Amplification primers were selected from the published mitochondrial reference sequence (1) using the Primer3 program (Whitehead Institute, MIT, Cambridge, MA). To select primers to amplify long tracts of DNA, such as mtDNA, in overlapping segments we have created a program known as PCR-Overlap that works together with Primer3. PCR-Overlap (available at http://droog.mbt.washington.edu) automates the selection of

*To whom correspondence should be addressed. Tel: +1 206 685 7339; Fax: +1 206 685 7301; Email: mrieder@u.washington.edu

primer sets based on the product size and product overlap defined by the user for the reference sequence of interest, e.g. the mitochondrial genome. Following selection by PCR-Overlap, each primer was checked against Genbank to ensure specificity for the mitochondrial genome using Blastn (http://www.ncbi.nlm.nih.gov) and scanned for the presence of known variants (MITOMAP, http://www.gen.emory.edu/mitomap.html) near the 3′-end of the primer which could influence priming efficiency. Prior to synthesis of the final set either a universal forward (–21 M13, TGT AAA ACG ACG GCC AGT) or reverse (M13reverse, CAG GAA ACA GCT ATG ACC) sequence (designated F and R in Table 1) was added to the 5′-end of each mitochondrial primer to produce PCR fragments compatible with dye–primer fluorescence-based sequencing. All primers were synthesized using standard phosphoramidite chemistry on an ABI 394 DNA synthesizer.

DNA amplification

mtDNA from 12 Caucasian individuals from CEPH (Centre d'etude Polymorphisme Humane) reference families were amplified in 20 µl reactions containing a standard PCR buffer [10 mM Tris–HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, 40 mM dNTPs, 0.5 mM primer, 0.5 U Taq polymerase (Perkin-Elmer Cetus, Norwalk, CT)] and 20 ng genomic DNA. The entire mitochondrial genome was amplified in 24 separate reactions using a single set of cycling conditions. Thermal cycling (GeneAmp PCR System 9600; Perkin-Elmer) was performed with an initial denaturation at 94°C for 1 min followed by 35 cycles of denaturation at 94°C for 30 s, primer annealing for 45 s at 61°C and primer extension at 72°C for 2 min. Following this a final extension was carried out at 72°C for 5 min.

DNA sequencing

Following DNA amplification unincorporated PCR primers and deoxynucleotide triphosphates in the samples were inactivated prior to sequencing by an enzymatic treatment. This was accomplished by mixing 6 µl PCR product with 1 µl exonuclease (10 U/µl; Amersham Life Science Inc., Arlington Heights, IL) and 1 µl shrimp alkaline phosphatase (2 U/µl; Amersham) and incubating at 37°C for 15 min followed by 80°C for 15 min to inactivate the exonuclease and alkaline phosphatase enzymes prior to sequencing. For sequencing the enzyme-treated PCR sample was subdivided into four separate reactions as follows: 1 µl each of the PCR sample mixed with 4 µl PRISM ready premix for the A and C reactions and 2 µl each of the PCR sample mixed with 8 µl PRISM ready premix for the G and T reactions (ABI PRISM Dye Primer Sequencing Kits with Amplitaq DNA polymerase FS; Perkin-Elmer Corporation, Foster City, CA). Sequencing reactions were denatured for 1 min at 96°C followed by 15 cycles at 96°C for 10 s, 55°C for 5 s and 70°C for 1 min and 15 cycles at 96°C for 10 s and 70°C for 1 min. After sequencing the A, C, G and T reactions were pooled and subjected to ethanol precipitation. The extension products were then evaporated to dryness under negative pressure (Savant Instruments, Farmingdale, NY), resuspended in 2.4 µl loading buffer (5:1, 1% deionized formamide/50 mM EDTA, pH 8.0), heated for 2 min at 90°C and loaded onto an Applied Biosystems 373A Stretch Sequencer.

Table 1. PCR primers used for mtDNA amplification

Primer Name[a]	Primer Sequence 5'-3'	3' Position[b]	Length	Overlap[c]
1F	CTCCTCAAAGCAATACACTG	611		
1R	TGCTAAATCCACCTTCGACC	1411	840	202
2F	CGATCAACCTCACCACCTCT	1245		
2R	TGGACAACCAGCTATCACCA	2007	802	204
3F	GGACTAACCCCTATACCTTCTGC	1854		
3R	GGCAGGTCAATTTCACTGGT	2669	860	196
4F	AAATCTTACCCCGCCTGTTT	2499		
4R	AGGAATGCCATTGCGATTAG	3346	887	208
5F	TACTTCACAAAGCGCCTTCC	3169		
5R	ATGAAGAATAGGGCGAAGGG	3961	832	215
6F	TGGCTCCTTTAACCTCTCCA	3796		
6R	AAGGATTATGGATGCGGTTG	4654	898	203
7F	ACTAATTAATCCCCTGGCCC	4485		
7R	CCTGGGGTGGGTTTTGTATG	5420	975	207
8F	CTAACCGGCTTTTTGCCC	5255		
8R	ACCTAGAAGGTTGCCTGGCT	6031	814	201
9F	GAGGCCTAACCCCTGTCTTT	5855		
9R	ATTCCGAAGCCTGGTAGGAT	6642	827	214
10F	CTCTTCGTCTGATCCGTCCT	6469		
10R	AGCGAAGGCTTCTCAAATCA	7315	886	211
11F	ACGCCAAAATCCATTTCACT	7148		
11R	CGGGAATTGCATCTGTTTTT	8095	987	205
12F	ACGAGTACACCGACTACGGC	7937		
12R	TGGGTGGTTGGTGTAAATGA	8797	900	196
13F	TTTCCCCCTCTATTGATCCC	8621		
13R	GTGGCCTTGGTATGTGCTTT	9397	816	214
14F	CCCACCAATCACATGCCTAT	9230		
14R	TGTAGCCGTTGAGTTGTGGT	10130	940	205
15F	TCTCCATCTATTGATGAGGGTCT	9989		
15R	AATTAGGCTGTGGGTGGTTG	10837	891	182
16F	GCCATACTAGTCTTTGCCGC	10672		
16R	TTGAGAATGAGTGTGAGGCG	11472	840	203
17F	TCACTCTCACTGCCCAAGAA	11314		
17R	GGAGAATGGGGGATAGGTGT	12076	802	196
18F	TATCACTCTCCTACTTACAG	11948		
18R	AGAAGGTTATAATTCCTACG	12772	866	166
19F	AAACAACCCAGCTCTCCCTAA	12571		
19R	TCGATGATGTGGTCTTTGGA	13507	977	242
20F	ACATCTGTACCCACGCCTTC	13338		
20R	AGAGGGGTCAGGGTTCATTC	14268	970	207
21F	GCATAATTAAACTTTACTTC	14000		
21R	AGAATATTGAGGCGCCATTG	14998	938	206
22F	TGAAACTTCGGCTCACTCCT	14856		
22R	AGCTTTGGGTGCTAATGGTG	15978	1162	180
23F	TCATTGGACAAGTAGCATCC	15811		
23R	GAGTGGTTAATAGGGTGATAG	5	765	205
24F	CACCATTCTCCGTGAAATCA	16420		
24R	AGGCTAAGCGTTTTGAGCTG	775	954	203

[a]All primers designated F or R were synthesized with the universal M13 (–21) forward primer or reverse primers (TGT AAA ACG ACG GCC AGT) or M13 reverse (CAG GAA ACA GCT ATG ACC) respectively on the 5′-end.
[b]Position of 3′-end of primer (1).
[c]Overlap (including primers) with preceding PCR fragment.

DNA sequence analysis and variation identification

The ABI sequence software (v.2.1.2) was used for lane tracking and first pass analysis (Perkin-Elmer Corporation). Chromatogram files were transferred to a Unix workstation (Sun Microsystems Inc., Mountain View, CA) and re-analyzed using the base calling software Phred (v.0.961028, available at http://www.genome.washington.edu). Phred provides improved base calling and a measure of data quality (range 0–50) for each base in a trace. Sequence quality can be linked to an error probability, i.e. accuracy of each base call (P.Green, personal communication). Three criteria are used to generate quality measures in Phred, including peak spacing, the relative size of the uncalled and called peaks and the change in signal between called peaks (P.Green and B.Ewing, personal communication).

Assembly of the sequencing reads and generation of the consensus genome sequence for each individual was performed using the program Phrap (v.0.960731). Phrap uses the base calls and quality information obtained from Phred and aligns each of the overlapping reads. Phrap uses sequence quality to generate a consensus sequence from the highest quality base calls in the final assembly and generates an adjusted quality (range 0–90) at each position by taking sequence context quality and opposite strand

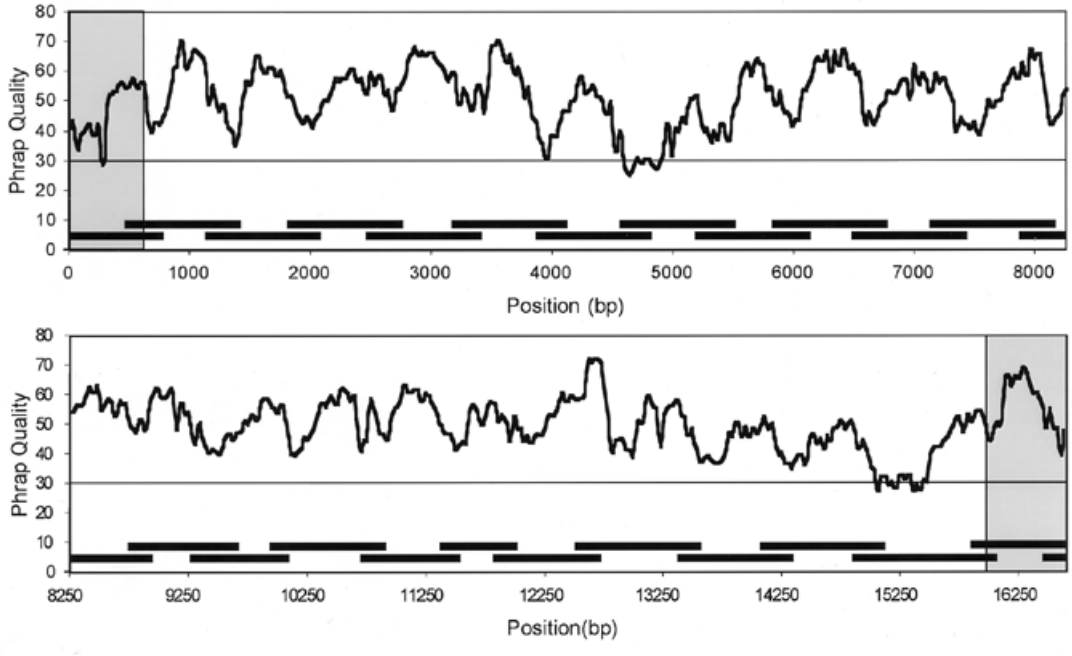

Figure 1. Average quality profile across the entire mitochondrial genome. The black line represents the average Phrap quality after final assembly of the consensus sequence from all 12 individuals. The bold lines underneath the graph show the overlapping PCR fragments covering each region. The average minimum acceptable quality (Phrap quality = 30) for each fragment is shown across the entire genome. Note that in the middle portion of many fragments the quality was adjusted upward to reflect the sequence confirmation from each strand. The mitochondrial control region is highlighted in grey.

confirmation into account (P.Green, personal communication). To simplify data analysis of each mitochondrial sequence three annotated versions of the Anderson reference sequence containing the locations of (i) the PCR primers, (ii) the coding and non-coding regions and (iii) the known mitochondrial variants (20) were also included in the Phrap assembly.

Potential DNA variants were automatically identified using two different programs: (i) RefComp, a program that compares a known reference sequence such as the Anderson sequence (1) with the consensus sequence for each individual and 'tags' nucleotide mismatches between the two sequences as potential variants for review; (ii) PolyPhred (21), a program that identifies heterozygous single nucleotide variants in fluorescence-based sequencing traces (both programs available at http://droog.mbt.washington.edu). In this study PolyPhred was used to automatically detect high levels of heteroplasmy among the 12 mitochondrial sequences. PolyPhred's accuracy in identifying heterozygotes using single pass sequences is >99% when the sequences are generated with the dye–primer chemistry and is related to the sequencing quality, which is decreased when dye–terminator chemistries are applied (21). All potential sites of variation tagged by either the RefComp or PolyPhred programs were reviewed by an analyst using the program Consed (v.4.1). During this initial review each potential site was classified as a true positive (high quality mismatch), a false positive site resulting from low sequence quality or, less frequently, as a base calling discrepancy.

RESULTS

Primer selection

In automatically selecting primers to amplify the mitochondrial genome we constrained the program PCR-Overlap to maximize the size of the amplified segment over a range of 800–1000 bp. This fragment length was selected because it is easily covered using just two sequencing reactions, i.e. a forward and reverse reaction, with some expected amount of overlap in the middle. We set the range for sequence overlap between PCR segments to be 150–200 bp. This ensured that the primer sequences used to amplify one fragment were scanned for the presence of variations by the segments that overlap the 5′- and 3′-ends. A final primer set was derived (Table 1) that amplified the entire human mitochondrial genome in 24 overlapping segments using a single set of cycling conditions. The average size of the amplified fragments was 892 bp (± 86 bp) and the average overlap of one fragment to the next was 203 bp (± 14 bp).

Sequence quality

To ensure that each PCR fragment was sequenced to produce the highest quality and fewest ambiguous or miscalled bases, universal primer sequences were added to the 5′-ends of each mitochondrial primer set. This chemistry produces higher quality sequences with longer reads and more uniform peak areas (21,22). Once the 24 amplified segments from an individual were sequenced from the forward and reverse directions these reads (48 altogether), comprising the complete genome sequence, were base called with Phred and assembled with Phrap. We have generated an average Phrap quality profile across the 12 individuals sequenced in this study (Fig. 1). The minimum acceptable average Phrap sequence quality across the genome was 30. This corresponds to an approximate error probability in base calling of 1 in 1000 (P.Green, personal communication). Furthermore, in many regions of the mitochondrial genome much higher qualities (quality >50, or an estimated error probability in base calling of 1 in 100 000) were achieved. The variability of

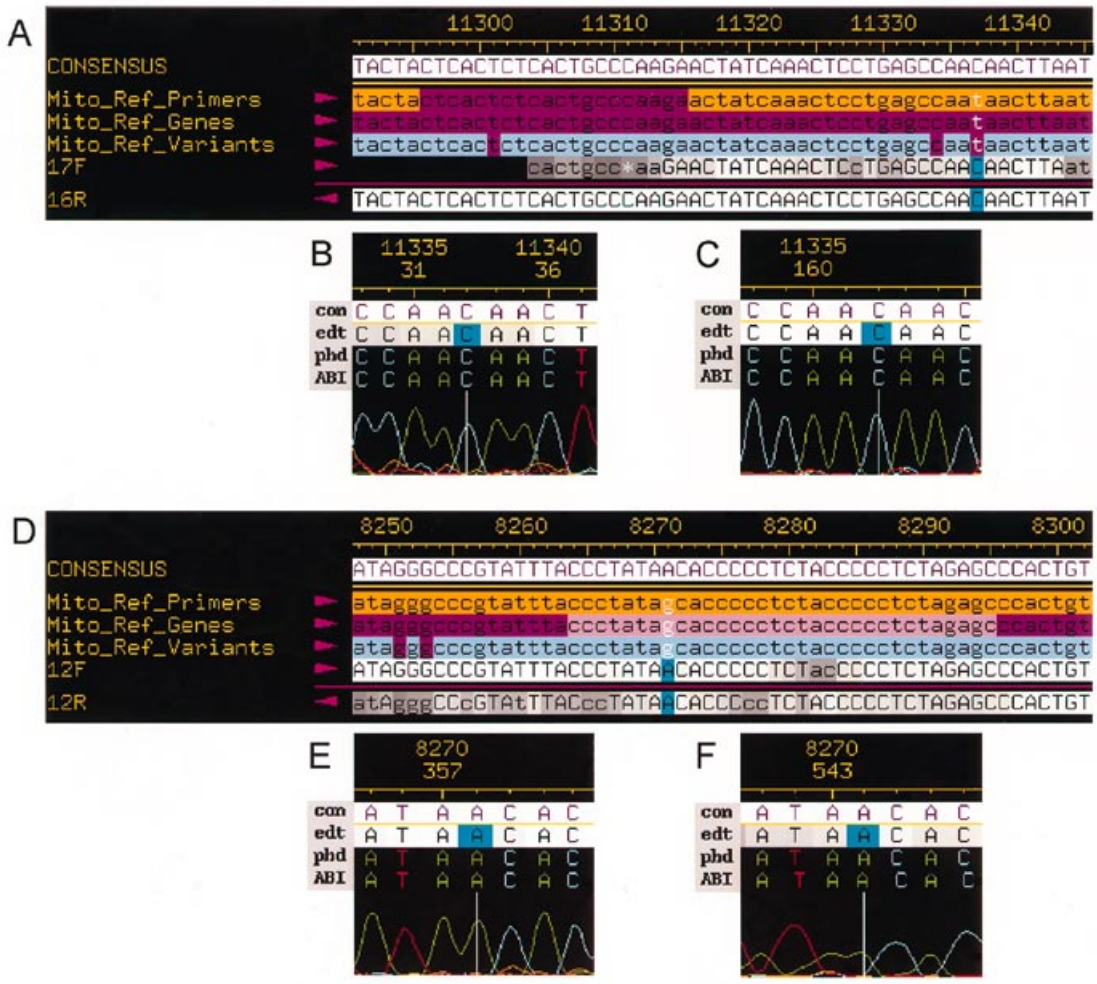

Figure 2. Assembly viewing, variant confirmation and sequence annotation using Consed. (**A** and **D**) Consed windows showing the consensus sequence constructed from the sequence reads (CONSENSUS), the annotated reference sequences for PCR primer locations (Mito_Ref_Primers, orange), gene regions (Mito_Ref_Genes, pink) and known variant sites (Mito_Ref_Variants, blue) highlighted with a purple tag and the aligned sequencing reads for an individual (F, forward; R, reverse). The sequence quality of the CONSENSUS and individual sequencing reads are encoded by a gray scale where gray and white represent low quality and high quality respectively. Putative variant sites were detected by identifying positions where the consensus sequence differed from the reference sequence (Mito_Ref_Variants) and were tagged (blue) by the RefComp program. (**A**) A previously known variant site at position 11337 T→C (11335 Anderson) in the NADH dehydrogenase 4 gene (**B** and **C**) sequencing traces showing confirmation of the 11335 variant site. (**D**) A novel variant site at position 8271 G→A (8269 Anderson) with (**E**)and (**F**) showing sequencing traces for the 8269 variant site confirmed on both strands.

Phrap quality across the mitochondrial genome is dependent on the amount of opposite strand sequence confirmation and typically the highest quality occurs near the middle of each fragment, where the forward and reverse sequencing reactions overlap. Even though overlaps are located at the beginning and end of adjacent fragments, these regions may have a lower sequence quality due to the reduced signal intensity of the overlapping strand and the presence of 'primer peaks'. There were three segments where the sequences on average yielded Phrap qualities near 30 (fragments 6, 7 and 22, Fig. 1). In contrast, one of the most sequenced regions in the human mitochondrion is the control region (positions 16 024–576, Fig. 1, shaded region), which yielded an average quality of 51 and, therefore, can be sequenced with high accuracy. Overall, dye–primer sequencing of the 24 PCR products generated reads with an average Phred quality of 33 and average length of 483 bp, producing ~1.4-fold coverage per genome.

Variant detection/use of annotated sequences

Potential variant sites in the mitochondrial sequences were identified by scanning for nucleotide mismatches between the mitochondrial reference sequence (1) and the base calls obtained by Phred for the 12 individuals sequenced. Each mismatched site was 'tagged' by the RefComp program. The 'tagging' process simplifies the subsequent review and analysis of these positions in the mitochondrial sequence using the Consed program. Once a site is 'tagged', navigation tools in Consed can be used to automatically move to each of these positions within the assembled sequences. These sites are also highlighted blue for easy identification of their location in the sequence reads [Fig. 2A (17F and 16R sequences) and D (12F and 12R sequences)].

In most large scale projects Consed is only used to view and edit the consensus sequence and the assembled reads (which are highlighted on a grey scale indicating the quality of the base call

Table 2. Novel variants and heteroplasmic sites detected in the mitochondrial genome

Position[a]	Base change[b]	Gene[c]	Amino Acid change
477	T→C	CTRL	--
513	G→A	CTRL	--
549	C→T	CTRL	--
575	C→T	CTRL	--
951	G→A	12S rRNA	--
1189	T→C	12S rRNA	--
1888	G→A	16S rRNA	--
2623H	G→A	16S rRNA	--
4295	A→G	$tRNA^{Ile}$	--
4639	T→C	ND2	Ile→Thr
6260	G→A	COI	--
7789	G→A	COII	--
7906	C→T	COII	--
8269	G→A	Non-Coding	--
8448	T→C	ATPase8	Met→Thr
8697	G→A	ATPase6	--
8898	C→T	ATPase6	--
8989	G→A	ATPase6	Ala→Thr
9150	A→G	ATPase6	--
9214	A→G	COIII	His→Arg
9316	T→C	COIII	Phe→Ser
10499	A→G	ND4	--
10993	G→A	ND4	--
13117	A→G	ND5	--
13401	T→C	ND5	--
13740	T→C	ND5	--
13759	G→A	ND5	Ala→Thr
14133	A→G	ND5	--
14386	T→C	ND6	--
15833	C→T	Cytb	--
16266H	C→T	CTRL	--

[a]Position in the Anderson sequence (1), (H), heteroplasmic.
[b]Base changes given as refence→variant.
[c]Gene name abbreviations: CTRL, control region; ND, NADH dehydrogenase; CO, cytochrome oxidase; ATPase, ATP synthase; Cytb, cytochrome b.

at each position). However, we have leveraged this program's capabilities to further simplify the process of data analysis by importing annotated reference sequences 'tagged' with known sequence features. As shown in Figure 2A and D, in addition to the consensus sequence and sequence reads from the individual under analysis, three annotated reference sequences are also present in the assembly. Each of these sequences stores specific information concerning the location of the: (i) PCR primers (Mito_Ref_Primers); (ii) mitochondrial genes (Mito_Ref_Genes); (iii) previously known mtDNA variants (Mito_Ref_Variants). When a feature is associated with a given sequence it is highlighted with purple and it has a specific window associated with this tag where information about the feature is stored, i.e. the specific PCR primer number, the gene, the nature of the known variant and reference(s) associated with its identification. This information can be retrieved within the Consed program. Therefore, when potential variants are viewed by the analyst (Fig. 2B and C and E and F) it is immediately clear where the variant is located with regard to the PCR primer, whether (Fig. 2A) or not (Fig. 2D) it is located in a gene segment and whether it is a previously known variant (Fig. 2A) or a novel variant (Fig. 2D).

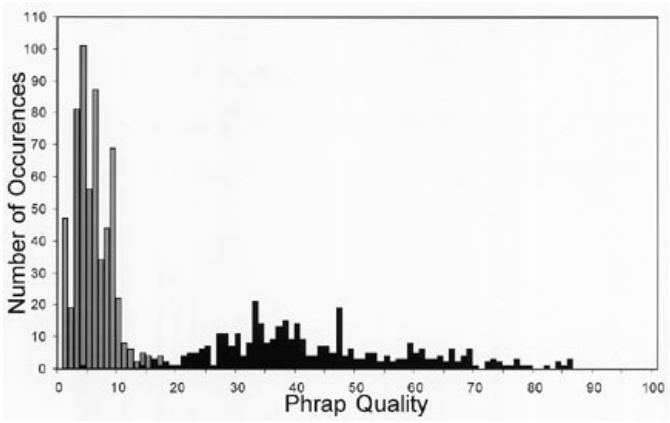

Figure 3. Distribution of true positive and false positive sites versus Phrap sequence quality. All variant sites were examined and classified as true positive (black bars) or false positive (gray bars) by an analyst and are plotted with respect to Phrap sequence quality at that site.

Mitochondrial variants and verification

The level of sequence quality obtained by fluorescence-based re-sequencing significantly improves the accuracy and sensitivity of automatically calling DNA variations among the assembled sequences. The program RefComp was used initially to identify all potential variant sites which mismatched the mitochondrial reference sequence regardless of sequence quality. Of the >278 359 bp screened in this study (12 individuals × 16 569 bp × 1.4-fold coverage), only 1054 sites (0.35% of all sites) were identified as potential variants (~87 sites/individual) in this first pass. Each of these potential sites was visually inspected using the Consed program (Fig. 2B and C and E and F) and classified as a true variant (378 sites, 35.9% of all potential variant sites), a false positive, which consisted of inconclusive or N base calls (597 sites, 56.6% of all potential variant sites) or as an incorrect base call (79 sites, 7.5% of all potential variant sites). Sequencing confirmation of variants could in many cases be obtained at extended read lengths (>500 bp, for example Fig. 2F). Of the variant sites which were classified as incorrect base calls, three sites were found consistently to be miscalled in all 12 individuals as the result of G-C compressions (reference sequence positions 85, 3959 and 15 003).

Further analysis of the distribution of the true positive variants and first pass false positive sites based on sequence quality (Fig. 3) revealed a clear distinction between these calls with respect to Phrap quality. We find that the majority of false positive sites are located in lower quality sequence. By setting a quality filter on the RefComp program and using a threshold of Phrap quality of ≥20 we found that 97.6% (369/378 sites) of the true positives were automatically identified, while accepting only 0.59% (4/679 sites) of the false positive sites. At this quality threshold an average of 95.1% of the mitochondrial genome (15 757/16 569 bp) was automatically scanned for variants with >99.98% accuracy on a single pass.

A total of 378 variants were identified during this analysis and, similar to previous studies (23), we have found that transitions (75.6%) were more prevalent than transversions (24.4%) among the variations. Because mtDNA has been so well studied the majority

of the 378 variants were previously known and only 29 of these variants were novel (i.e. unreported in GenBank or MITOMAP; see Table 2). Of these novel variants 19 were found in gene coding regions, leading to six amino acid changes. Four variants were found in rRNAs, four in the control region, one in a tRNA and one in a non-coding region. All novel variants were unique to the individual where identified (i.e. not found in any other individual) and each was confirmed by opposite strand sequencing or re-amplification and re-sequencing of the individual at that site. Furthermore, we have found 14 positions that were consistently different from the Anderson reference sequence in all 12 of the genomes sequenced: 263:(G→A) (position in the reference sequence: nucleotide reported in the Anderson sequence→nucleotide detected in the 12 individuals), 750:(A→G), 3107:(C→$C_{deletion}$), 3423*:(G→T), 4985:(G→A), 8860:(A→G)*, 9559:(G→C), 11 335:(G→C), 13 702*:(G→C), 14 199*:(G→A), 14 272*:(G→C), 14 365*:(G→C), 14 368*:(G→C) and 15 326:(A→G). A subset of these sites (indicated by an asterisk) have previously been suggested to be invariant positions (3). Removal of these invariant sites from the total number of sites yielded 215 variant sites among the 12 individuals sequenced (18.0 variants/individual, range 5–30 variants/individual) and a mean pairwise difference of 19.1 sites/individual. Furthermore, the rate of DNA variation in the D loop was 8.5-fold higher than that of the other mitochondrial sequences.

Detection of heteroplasmy

We have also scanned the mitochondrial sequences for potential sites with high levels of heteroplasmy. To accomplish this we used PolyPhred, a program that detects the presence of heterozygous bases in nuclear DNA sequences (21). Specifically, we searched for levels of heteroplasmy which would be comparable with levels associated with a heterozygous site in nuclear DNA (i.e. 50/50% split). With PolyPhred we automatically detected two heteroplasmic positions in two different individuals, as shown in Figure 4. Furthermore, in each case the presence of heteroplasmy was verified by opposite strand sequencing. Based on the drop in the primary peak (24), we estimate the levels of heteroplasmy at position 2623 to be 40/60% (A/G) and at position 16 266 to be 60/40% (C/T).

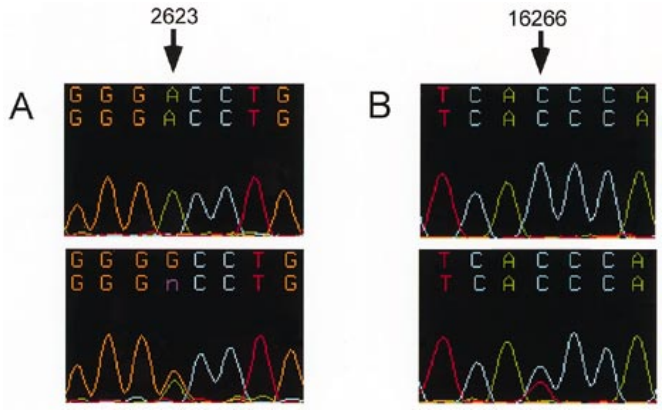

Figure 4. Detection of mitochondrial heteroplasmy using fluorescence-based DNA sequencing. Examples of substitution heteroplasmy (shown by the black arrow) found at positions (**A**) 2623 and (**B**) 16 266. The top panels in (A) and (B) show an individual homoplasmic at each position. The bottom panels shows the heteroplasmic site on the forward strand. Heteroplasmy was estimated to be C/T = 60/40% in (A) and A/G= 40/60% in (B) and was confirmed on the opposite strand.

DISCUSSION

There are many approaches available for identifying DNA variations in a sequence of interest. Most approaches are two tiered, where some form of comparative gel or hybridization analysis is carried out, followed by direct DNA sequencing to identify or confirm the nature and location of the variants identified during the primary scan (25,26). Because of its sensitivity and accuracy, fluorescence-based sequencing is already considered the 'gold' standard for detecting DNA variations (27). Furthermore, the large scale application of fluorescence-based sequencing in generating reference sequences for simple and complex genomes is continuing to drive the evolution of this technology, increasing its availability and decreasing its costs. For example, new sequencing chemistries based on fluorescence energy transfer are producing higher quality sequences with less DNA template and over time these chemistries will likely replace the standard chemistries currently in use (28–30). Additionally, the very high signal-to-noise ratios generated with these energy transfer chemistries may lead to the development of more efficient ways to identify single nucleotide variations using pooled sequencing templates (24).

In this report we have leveraged the use of new analysis tools in automating and simplifying the identification of DNA variations across long tracts of DNA. Two programs widely used in generating reference genome sequences are the Phred base caller and the Phrap assembly program. Each of these provides a quantitative measure of sequence quality that can be linked to an objective estimate of error rate in calling the final order of nucleotides in a sequence (P.Green, personal communication). The availability of sequence quality parameters is important for developing standards for acceptable sequence quality when performing diagnostic re-sequencing. Although all the detected mitochondrial variants were visually reviewed and classified by an analyst, a completely automated system for classifying sites based on sequence quality could be derived that would not require human intervention or decision making. Analysis of the distribution of sequence quality between true variant and false positive sites showed a clear demarcation based on sequence quality (Fig. 3), permitting one to set threshold limits for acceptable sequence analysis automatically. Setting a threshold limit of Phrap quality ≥20 would have excluded 99.42% of the false positive sites while accurately identifying 97.6% of the true positives. Because sequence quality can be related to a base calling error probability, the distribution of quality values across all individuals was used to estimate the theoretical number of sites that would be incorrectly called by Phred at this quality threshold and was calculated to be four false positive sites/individual. Our results indicate a false positive rate of 0.33 false negative sites/individual, which suggests that calculations used for error estimation by Phred may be slightly conservative. Although we did not directly attempt to identify false negative sites, these would occur when a site was analyzed by Phred and called the same base as the reference sequence but in fact was a variant site. The theoretical false negative rate was estimated to be 1.3 sites/individual and, based on the conservative nature of the error estimation by Phred for false positive sites, could be much lower. These results clearly illustrate the utility of

using quality measures as a criterion for sequence trace analysis and the need for a greater understanding of the relation between sequence quality and variant identification.

Another advantage of this analysis system is the interface between Phred and Phrap and the Consed program. Consed provided a uniform environment for viewing sequence data and simplified the localization of tagged DNA variants through its navigation system. Furthermore, because information can be stored in Consed tags, annotated sequences containing the existing information on a gene or sequence can be viewed along with the newly generated sequences (Fig. 2A–F). This greatly simplifies the subsequent analysis of a new variant in terms of the information available on the reference sequence. Furthermore, the exchange of an annotated common reference sequence with all known variant sites and significant features, between investigators and databases, would simplify variation analysis (23,31).

Numerous studies have recently suggested that mitochondrial heteroplasmy may be more prevalent than previously thought (32–34). The mtDNA is unique because it can be heteroplasmic, having two or more different mtDNA sequences within an individual (35) and also different levels of heteroplasmic mtDNA in different tissues (36). We have also applied an automated approach in the detection of high levels of heteroplasmy with fluorescence-based sequencing and found two sites in two different individuals. While similar fluorescent sequencing methods have been used to detect heteroplasmy (37,38), other more sensitive cloning methods have been used to detect lower levels of heteroplasmy. More automated estimates of heteroplasmy like those attempted in this study could provide additional insight into these questions.

In summary, this study demonstrates the feasibility of performing rapid, high quality, accurate fluorescence-based sequencing and DNA variant detection in the mitochondrial genome in a single pass. The use of a combined set of DNA sequence analysis programs, Phred, Phrap, Consed and PolyPhred, allowed for quantitative estimates of sequence quality, visual display of integrated sequence information and detection of heteroplasmy. This system will simplify analysis of the mitochondrial genome within and between individuals and could easily be extended to a system automating detection of single nucleotide variants in large tracts of nuclear DNA.

ACKNOWLEDGEMENTS

This work was supported in part by grants DIR8809710, HG01436 and DE-FG03-97ER-62385 to D.A.N. and a genome training grant fellowship to M.J.R. (HG00035). We thank Drs Phil Green and Brent Ewing and Mr David Gordon for sharing their insights and programs Phred, Phrap and Consed with us and Dr Andy Clark for his helpful comments. We also thank the developers of MITOMAP: the Mitochondrial Human Genome Database at Emory University in Atlanta (http://www.gen.emory.edu/mitomap.html) for information on previously known mitochondrial variants.

REFERENCES

1 Anderson,S. *et al*. (1981) *Nature*, **290**, 457–465 (Genbank accession no. J01415).
2 Wallace,D.C. (1992) *Science*, **256**, 628–632.
3 Horai,S., Hayasaka,K., Kondi,R., Tsugane,K. and Takahata,N. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 532–536.
4 Stoneking,M. and Soodyall,H. (1996) *Curr. Opin. Genet. Dev.*, **6**, 731–736.
5 Wilson,M.R., DiZinno,J.A., Polanskey,D., Replogle,J. and Budowle,B. (1995) *Int. J. Legal Med*., **108**, 68–74.
6 Sullivan,K.M., Hopgood,R. and Gill,P. (1992) *Int. J. Legal Med*., **105**, 83–86.
7 Piercy,R., Sullivan,K.M., Benson,N. and Gill,P. (1993*) Int. J. Legal Med*., **106**, 85–90.
8 Wallace,D.C., Shoffner,J.M., Trounce,I., Brown,M.D., Ballinger,S.W., Corral-Debrinski,M., Horton,T., Jun,A.S. and Lott,M.T. (1995) *Biochim. Biophys. Acta*, **1271**, 141–151.
9 Larsson,N.F. and Clayton,D.A. (1995) *Annu. Rev. Genet*., **29**, 151–178.
10 Horai,S., Murayam,K., Hayasaka,K., Matsubayashi,S., Hattori,Y., Fucharoen,G., Harihara,S., Park,K.S., Omoto,K. and Pan,I.H. (1996) *Am. J. Hum. Genet*., **59**, 579–590.
11 Sajantila,A. *et al*. (1995) *Genome Res*., **5**, 42–52.
12 Watson,E., Bauer,K., Aman,R., Weiss,G., Haeseler,A. and Paabo,S. (1996) *Am. J. Hum. Genet*., **59**, 437–444.
13 Cheng,S., Higuchi,R. and Stoneking,M. (1994) *Nature Genet*., **7**, 350.
14 Yoon,K.L., Modica-Napolitano,J.S., Ernst,S.G. and Aprille,J.R. (1991) *Anal. Biochem*., **196**, 427–432.
15 Cann,R.L., Stoneking,M. and Wilson,A.C. (1987) *Nature*, **325**, 31–36.
16 Jaksch,M., Gerbitz,K.D. and Kilger,C. (1995) *Clin. Biochem*., **28**, 503–509.
17 Stoneking,M., Hedgecock,D., Higuchi,R.G., Vigilant,L. and Erlich,H.A. (1991) *Am. J. Hum. Genet*., **48**, 370–382.
18 Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P. (1996) *Science*, **274**, 610–614.
19 Tully,G., Sullivan,K.M., Nixon,P., Stones,R.E. and Gill,P. (1996) *Genomics*, **34**, 107–113.
20 Center for Molecular Medicine (1995) *MITOMAP: the Human Mitochondrial Genome Database*. Center for Molecular Medicine, Emory University, Atlanta, GA, http://www.gen.emory.edu/mitomap.html.
21 Nickerson,D.A., Tobe,V.O. and Taylor,S.L. (1997) *Nucleic Acids Res*., **14**, 2745–2751.
22 Phelps,R.S., Chadwick,R.B., Conrad,M.P., Kronick,M.N. and Kamb,A. (1995) *BioTechniques*, **19**, 984–989.
23 Marzuki,S., Noer,A.S., Lertrit,P., Thyagarajan,D., Kapsa,R., Utthanaphol,P. and Byrne,E. (1991) *Hum. Genet*., **88**, 139–145.
24 Kwok,P., Carlson,C., Yager,T.D., Ankener,W. and Nickerson,D.A. (1994) *Genomics*, **23**, 138–144.
25 Thomas,A.W., Edwards,A., Sherratt,E.J., Majid,A., Gagg,J. and Alcolado,J.C. (1996) *J. Med. Genet*., **33**, 253–256.
26 Hacia,J.G., Brody,L.C., Chee,M.S., Fodor,S.P. and Collins,F.S. (1996) *Nature Genet*., **14**, 441–447.
27 Eng,C. and Vijg,J. (1997) *Nature Biotechnol*., **15**, 422–426.
28 Ju,J., Glazer,A.N. and Mathies,R.A. (1996) *Nature Med*., **2**, 246–249.
29 Metzker,M.L., Lu,J. and Gibbs,R.A. (1996) *Science*, **271**, 1420–1402.
30 Lee,L.G., Spurgeon,S.L., Heiner,C.R., Benson,S.C., Rosenblum,B.B., Menchen,S.M., Graham,R.J., Constantinescu,A., Upadhya,K.G. and Cassel,J.M. (1997) *Nucleic Acids Res*., **25**, 2816–2822.
31 Kogelnik,A.M., Lott,M.T., Brown,M.D., Navathe,S.B. and Wallace,D.C. (1996) *Nucleic Acids Res*., **24**, 177–179.
32 Stoneking,M. (1996) *Biol. Chem*., **377**, 603–604.
33 Comas,D., Pääbo,S. and Bertranpetit,J. (1995) *Genome Res*., **5**, 89–90.
34 Ivanov,P.L., Wadhams,M.J., Roby,R.K., Holland,M.M., Weedn,V.W. and Parsons,T.J. (1996) *Nature Genet*., **12**, 417–420.
35 Bidooki,S.K., Johnson,M.A., Chrzanowska-Lightowlers,Z., Bindoff,L.A. and Lightowlers,R.N. (1997) *Am. J. Hum. Genet*., **60**, 1430–1438.
36 Keightley,J.A., Hoffbuhr,K.C., Burton,M.D., Salas,V.M., Johnston,W.S.W., Penn,A.M.W., Buist,N.R.M. and Kennaway,N.G. (1996) *Nature Genet*., **12**, 410–415.
37 Gill,P., Ivanov,P.L., Kimpton,C., Piercy,R., Benson,N., Tully,G., Evett,I., Hagelberg,E. and Sullivan,K. (1994) *Nature Genet*., **6**, 130–135.
38 Parsons,T.J., Muniec,D.S., Sullivan,K., Woodyatt,N., Alliston-Greiner,R., Wilson,M.R., Berry,D.L., Holland,K.A., Weedn,V.W., Gill,P. and Holland,M.M. (1997) *Nature Genet*., **15**, 363–368.