Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq

Iouri Chepelev, Gang Wei, Qingsong Tang and Keji Zhao*

Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20892, USA

Received December 10, 2008; Revised May 22, 2009; Accepted May 26, 2009

ABSTRACT

Whole-genome resequencing is still a costly method to detect genetic mutations that lead to altered forms of proteins and may be associated with disease development. Since the majority of disease-related single nucleotide variations (SNVs) are found in protein-coding regions, we propose to identify SNVs in expressed exons of the human genome using the recently developed RNA-Seq technique. We identify 12176 and 10621 SNVs, respectively, in Jurkat T cells and CD4+ T cells from a healthy donor. Interestingly, our data show that one copy of the TAL-1 proto-oncogene has a point mutation in 3′ UTR and only the mutant allele is expressed in Jurkat cells. We provide a comprehensive dataset for further understanding the cancer biology of Jurkat cells. Our results indicate that this is a cost-effective and efficient strategy to systematically identify SNVs in the expressed regions of the human genome.

INTRODUCTION

Mutations in DNA sequence can alter or disrupt cellular function and lead to disease development through various pathways depending on their genomic location. Although mutations in noncoding regions may disrupt functional cis-regulatory elements that control transcription and lead to altered transcript levels, the majority of genetic diseases identified so far are linked to mutations in protein coding regions of the genome that lead to altered forms of proteins. Whole-genome resequencing of individual human genome at a high-enough coverage should be able to determine most of single nucleotide polymorphisms (SNPs) as well as somatic point mutations in coding and noncoding regions (1). However, it is very costly to do so at present. Since the majority of human genome mutations associated with diseases are either missense/nonsense or affecting splicing in only about 2% of the human genome (2), it would significantly reduce cost and increase efficiency by resequencing the exonic regions. We propose to identify point mutations in the expressed coding regions of the human genome by sequencing cDNAs using RNA-seq. This strategy can provide both the gene expression and single nucleotide variation (SNV) information at the same time. Since expressed regions constitute only a minor fraction of the human genome, sufficient coverage of these regions can be achieved at low sequencing cost using the next-generation sequencing technologies. Our data demonstrate that this approach efficiently identifies SNVs that alter protein sequence.

MATERIALS AND METHODS

Sample preparation and RNA sequencing

mRNAs isolated from Jurkat T cells, derived from acute lymphoblastic leukemia (ALL), and CD4+ T cells from a healthy donor, were converted to cDNA using standard protocols. The cDNAs were fragmented to 100–200 bp using sonication, followed by end repair and Solexa adapter ligation. The products were sequenced on an Illumina GAIi system according to established procedures (3–6).

Short reads alignment

The short reads were analyzed as outlined in Figure 1. The quality-filtered 30-bp short sequence reads were aligned to the reference sequence consisting of hg18 (NCBI Build 36) human genome plus a library of synthetic exon junction sequences using ELAND (Efficient Local Alignment of Nucleotide Data) software, allowing up to two mismatches with the reference sequence. There are three types of uniquely aligned reads: U0 reads that align perfectly to the reference sequence, U1 reads that have one mismatch with the reference sequence and U2 reads that have two mismatches with the reference sequence.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
About 25% of reads in our samples are U1 reads and 13% are U2 reads. The library of exon junction sequences was created as follows. Human exon sequences were retrieved from Ensembl database (release 50). We joined all possible pairs of exons that belong to the same transcript such that the genomic order of exons is respected. A junction sequence consists of last 26 bp of 5' exon and first 26 bp of 3' exon. Redundant sequences were removed from the resulting set of 52-bp exon junction sequences.

Redundant reads filtering

In order to remove possible PCR amplification artifacts and to reduce confounding effects of systematically bad sequencing cycles in short sequence reads, the following two consecutive filters were applied to the set of uniquely aligned reads (Figure 2A).

(i) Retain only a single copy of each read. (A read is defined as a string of letters A, C, G, T and N.)
(ii) There can still be multiple U1 and U2 reads that passed Filter 1 at the same genomic location. (Note that there can be at most one single U0 read that passed Filter 1 at each genomic location). Randomly select only one read each from U1 and U2 reads that map to the same location.

As a result of application of filters 1 and 2, at most three reads are retained at any given genomic location (bottom box). However, the stringent procedure of randomly selecting at most one read from U1 and U2 reads that map to the same location should not significantly reduce the statistical power to detect SNVs. As can be seen from the example in Figure 3A, there can be plenty of overlapping but noncoincident reads that cover an SNV. In fact, for reads of length 30 bp there can be as many as \( \frac{30}{2} = 90 \) reads that cover an SNV. Since 90-fold coverage is the upper-bound for coverage possible by filtered reads, the number of reads in very highly expressed exons will not correspond to actual expression levels. However, it should not be a concern because the

Figure 2. Redundant Reads Filter and SNV probability calculation examples. (A) There are nine reads that map uniquely to the same genomic location (top box). Nucleotide mismatches with reference sequence are highlighted in red. Filter 1 retains a single copy of each read. Thus, only five reads remain after Filter 1 is applied (middle box). There are two U1 reads, two U2 reads and one U0 read in the middle box. Filter 2 randomly selects one U1, one U2 and one U0 read. This leaves three reads at the same genomic location (bottom box). (B) Example of SNV probability calculation. Colored in red is a candidate SNV site. Seven short reads map uniquely to that site. The reference nucleotide is T. Five reads have nucleotides that differ from the reference nucleotide and two reads have nucleotide T at the candidate SNV site. Let the error rate estimated from the total number of U0, U1 and U2 nonredundant reads be \( q = 0.02 \). The binomial (random chance) probability to observe two matches and five mismatches at the same location is proportional to \( q^5 \left(1-q\right)^2 \). The \( P \)-value is given by the binomial probability of observing five or more mismatches in a seven-read alignment and it is equal to \( 6.5 \times 10^{-8} \).
we compared the numbers of known and unknown SNVs mapping but noncoincident reads. The evidence for presence of SNV comes mainly from overlapping but noncoincident reads. There are many overlapping but noncoincident reads that can cover a single SNV. In fact, there can still be as many as 90 reads of length 30 bp that cover a single SNV after the filtering step. Thus, the statistical power to detect the SNV is not reduced by the filtering procedure. (B) The number of detected \((P\text{-value} = 10^{-9})\) known, i.e. SNPs from dbSNP database, and unknown (novel) SNVs using reads filtered using four different filters: Filter A is the Redundant Reads Filter; Filter B is Filter 1 followed by randomly selecting two reads each from U1 and U2 categories; Filter C is Filter 1 followed by randomly selecting three reads each from U1 and U2 categories; the last filter is an empty filter, i.e. no filtering of unique reads is done. The number of detected known SNVs is not sensitive to the filtering method used, confirming very low false-positive rate among detected known SNVs. However, the number of detected unknown SNVs is much higher for the cases of Filters B, C and No filter than for Filter A, demonstrating high false-positive rates resulting from the use of these alternative filters. Thus, Filter A is the best of four filters.

The purpose of the filtering procedure is to reduce false positive rate of SNV detection and 90-fold coverage is a very significant coverage. By restricting the number of reads that can map to the same genomic location, we allow the evidence for presence of SNV come mainly from overlapping but noncoincident reads.

In order to further justify our use of Filter 1 + Filter 2, we compared the numbers of known and unknown SNVs detected using the following four different filtering procedures:

(a) Apply Filter 1 and then randomly select one read each from U1 and U2 categories. This is our Filter 1 + Filter 2.
(b) Apply Filter 1 and then randomly select two reads each from U1 and U2 categories.
(c) Apply Filter 1 and then randomly select three reads each from U1 and U2 categories.
(d) No filter is applied.

We applied these four filters to unique reads from CD4\(^+\) sample. The results shown in Figure 3B suggest that Filters B–D lead to significant increase in the number of false positives while only marginally increasing number of true positives. Let us argue that the last statement is indeed true. Exons expressed in CD4\(^+\) cells contain SNVs. Some of these SNVs are known SNVs, i.e. previously annotated in dbSNP database, while some SNVs are novel, i.e. previously unknown SNVs. Let the ratio of number of known to known SNVs in expressed exons be \(U:K\). An SNV-detection algorithm blind to the distinction between known and unknown SNVs will detect a certain number of SNVs. Let \(K_d\) and \(U_d\) be numbers of detected known and unknown SNVs, respectively. We expect that the ratio \(U_d/K_d\) is comparable to \(U/K\). The fact that \(U_d/K_d\) ratio for SNVs detected following the filters B, C or D is significantly higher than the corresponding ratio for filter A implies that a large fraction of novel SNVs detected following the filters B–D are actually false positives. Note that false-positive rate among detected known SNVs is low since the probability of sequencing errors occurring at precisely dbSNP SNP locations with the right alternative allele is small. This is supported by the fact that the number of detected known SNVs is not very sensitive to the filtering procedure used (see Figure 3B).

Assigning statistical significance to candidate SNV sites

The overall error rate of sequencing for each sample is estimated as the frequency of nucleotide mismatches. Using numbers of matches and mismatches from Table 1, the sequencing error rates for Jurkat and CD4\(^+\) samples are estimated to be 0.017 and 0.019, respectively. We use a simple binomial distribution to model null distribution of number of mismatches at a genomic locus with \(N\) uniquely aligned reads (Figure 2B). This model assumes that error rate is independent of the position in the read. It should be straightforward to extend our model to a model with position-dependent error rate.

Since there are a large number of candidate variant sites, nominal \(P\)-values calculated using the binomial null distribution need to be adjusted for multiple statistical testing. We used a very conservative multiple-testing procedure—Bonferroni method—to adjust nominal \(P\)-values. With around 6 million candidate variant sites, Bonferroni-adjusted \(P\)-value 0.01 corresponds to the nominal \(P\)-value \(0.01/6\times 10^{6} = 1.7 \times 10^{-9}\). In this paper, we called a variant site significant if the nominal
Table 1. Miscellaneous information about short sequence reads and single nucleotide variant sites detected in Jurkat and CD4⁺ T cell samples using RNA-seq data

|                        | Jurkat       | CD4        |
|------------------------|--------------|------------|
| Number of unique genomic reads | 26275213     | 27253287   |
| Number of unique exon-junction reads | 2451140      | 2267059    |
| Number of unique and nonredundant genomic reads | 13166074      | 13325274   |
| Number of unique and nonredundant junction reads | 1166914      | 924220     |
| Number of matched bases | 422487847    | 419253929  |
| Number of mismatched bases | 7501793      | 8230891    |
| Number of variant sites | 5477131      | 6030145    |
| Number of significant variant sites | 12176       | 10621     |
| Number of novel significant variant sites | 4703        | 2952      |
| Number of nonsynonymous mutations | 3206      | 1977      |
| Number of novel nonsynonymous mutations | 1770       | 747       |
| Number of nonsense mutations | 47        | 17       |
| Number of novel nonsense mutations | 41         | 9       |
| Number of splice-site mutations | 66         | 31       |
| Number of novel splice-site mutations | 60         | 23       |

Miscellaneous information about short sequence reads and single nucleotide variant sites detected in Jurkat and CD4⁺ T cell samples using RNA-seq data.

The number of unique and nonredundant reads refers to the number of uniquely mapped reads that pass Redundant Reads Filter. The number of mismatched bases refers to the total number of bases in unique nonredundant reads that mismatch with the reference sequence; it is given by the formula: number of U1 reads + twice the number of U2 reads.

The number of variant sites refers to the total number of genomic locations where mismatches with short sequence reads were observed. ‘Significant’ means significant at *P*-value < 1.0 × 10⁻⁹. ‘Novel’ means not present in dbSNP build 126 database.

*P*-value is <1.0 × 10⁻⁹. There are less conservative multiple-testing procedures available such as Benjamini–Hochberg method. These less conservative methods should be used to test statistical significance of candidate SNV sites with low reads coverage. In addition, the position dependence of error rates in short reads cannot be ignored for SNV sites with low reads coverage.

**Cost analysis for SNV detection**

The theoretical upper bound for the number of SNVs detectable using cDNA of expressed exons in a tissue is given by the number of expressed exons of the tissue. At a given sequencing depth *D*, a fraction *F* of these expressed exons will be detected. We estimate the curve *F(D)* as follows. Let *N* be the total number of nonredundant unique 30-bp sequenced reads.

In order to avoid the problem of double counting in regions where exons overlap, we merge all Ensembl exons and consider merged exonic regions. For each merged exonic region *E_k*, we compute reads per kilobase of exon model per million mapped reads (3) RPKM value *R_k* of its expression as *R_k* = (*10⁶ × N_k*)/(*N × L_k*), where *N_k* is the number of nonredundant unique reads in *E_k* and *L_k* is the length of *E_k*. At the sequencing depth we achieved (400 Mbp ~13 million nonredundant unique 30-bp genomic reads in CD4⁺ sample), we compute *R_k* for all merged exonic regions. We then assume that *R_k* value is independent of sequencing depth. We validated the latter assumption by comparison of *R_k* values computed using full CD4⁺ data set and 50% of the dataset (= four lanes of Solexa data). The Pearson correlation coefficient for this comparison is found to be >0.99.

Given the sequencing depth *D*, exon expression level *R_k* can be used to estimate coverage *C_k* of the exon by sequence reads. At the sequencing depth *D*, the expected number of 30-bp nonredundant unique reads in the exon is *N_k* = (*R_k* × *D* × *L_k*)/(*30 × 10⁹*). Thus, the expected coverage at sequencing depth *D* is given by *C_k* = (*30 × N_k*)/(*L_k* × *D*)/(*10⁹*).

The stringent *P*-value cutoff 10⁻⁹ used in this paper is related to fold coverage *C* as follows. The *P*-value for a heterozygous SNV with wt to mutant reads ratio 7:7 is ~10⁻⁹ when sequencing error rate is *q* = 0.019 (see example in Figure 2B). Thus, our *P*-value cutoff 10⁻⁹ corresponds to the fold coverage threshold *C* = 14. For sequencing depth that we achieved, *D* = 400 Mbp, coverage fold *C* = 14 corresponds to RPKM value ~35. Thus, heterozygous SNVs in exonic regions with RPKM <35 are undetectable at the sequencing depth *D* = 400 Mbp and *P*-value cutoff 10⁻⁹. It is easier to detect homozygous SNVs. At the sequencing error rate *q* = 0.019, homozygous SNV with wt to mutant reads ratio 0:5 has *P*-value ~10⁻⁹. The coverage *C* = 5 corresponds to RPKM value ~13.

With a large enough sequencing depth, most SNVs in expressed exons will eventually be detected. It has been shown in (3) that RPKM value 1 corresponds approximately to one transcript per cell. We thus assume that an exonic region *E_k* is expressed if its RPKM value *R_k* ≥ 1. The total size of merged exonic regions with RPKM ≥ 1 is ~35 Mbp which is about 50% of all merged exonic regions (~69 Mbp) in the human genome. Since SNVs should be more or less uniformly distributed in exonic regions, we introduce the density *ρ* of SNVs per base pair of exonic region. The expected number of SNVs in the exonic region *R_k* is thus equal to *ρ × L_k*. Given a SNV detection fold coverage threshold *C*, we assume that all SNVs in an exonic region *R_k* are detected once the fold coverage *C_k* of the region exceeds the threshold *C*. As the sequencing depth increases, more and more exonic regions pass coverage threshold. The expected fraction *F(D)* of expressed SNVs detected at the sequencing depth *D* is given by the equation:

\[ F(D) = (1/Z)\Sigma_k \Theta(C_k > C) \times L_k \times \Theta(R_k \geq 1), \]

where *C_k* = (*R_k* × *D*)/(*10⁹*), *Z* = *Σ_k L_k × Θ(R_k ≥ 1) is the total size (in bp) of expressed merged exonic regions and the function Θ of the Boolean variable is given by Θ(true) = 1, Θ(false) = 0. In this manner, we derived the SNV detection cost curve 100 × *F(D)* shown in Figure 4B.

**Exon sequencing coverage analysis**

The coverage analysis results shown in Figure 4A were obtained as follows. In order to avoid double-counting sequencing tags in the regions where exons overlap, we merged all exons and considered merged exonic regions. For each contiguous exonic region *E_k* we compute
sequencing coverage as $C_k = (30 \times N_k)/L_k$, where $N_k$ is the number of unique and non-redundant 30-bp sequencing reads detected in the region and $L_k$ is the length of the region. The fraction $P$ of exonic regions passing sequencing coverage threshold $C$ is defined as the ratio of the total size of exonic regions covered at least $C$-fold to the total size of all exonic regions: $P(C) = (1/L) \sum_k \Theta(C_k \geq C) \times L_k$, where $L = \sum_k L_k$.

**Annotation of significant SNV sites**

Genome sequence files, and Ensembl and RefSeq gene annotations tables for hg18 human genome were downloaded from UCSC Genome Bioinformatics website: http://genome.ucsc.edu. We determined synonymous/non-synonymous status of SNVs based on the codon changes resulting from nucleotide substitutions. An SNV was annotated to be a splice-site SNV if it lies in the first or last 2 bp of an intron.

**Software and data set**

We implemented Redundant Reads Filter and Point Mutation Analyzer in UNIX shell, Perl and C++. The entire suite of software is available at: http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/pma. On the Dell Precision T7400 Linux workstation, the analysis of 27 million unique reads from CD4+ sample took only 12 min: 8 min at filtering step using Redundant Reads Filter and 4 min for assembly of reads, SNV detection and assigning significance $P$-values to SNVs using Point Mutation Analyzer. As an input, the algorithm uses ELAND results files from Illumina Genome Analyzer. We deposited raw and processed data set for Jurkat and CD4+ T cells to Gene Expression Omnibus (GEO) database under accession number GSE16190.

**RESULTS AND DISCUSSIONS**

**Reads mapping, redundancy filtering and SNV detection**

In order to detect SNVs in the protein-coding and untranslated regions of the human genome using the next generation sequencing techniques, we designed a strategy as outlined in Figure 1. cDNAs synthesized from mRNAs were fragmented to 100–200 bp by sonication and sequenced using Illumina Genome Analyzer II. The short reads of 30 bp were mapped to the reference consisting of hg18 human genome plus a collection of synthetic exon junctions using ELAND software, allowing up to two mismatches with the reference (see ‘Materials and methods’ section).

The mismatches with the reference sequence can occur due to sequencing errors or point mutations present in the sample. In order to distinguish between these two possibilities and hence filter noise from signal, we applied the following two-step procedure to the set of uniquely mapped reads (see ‘Materials and methods’ section). Multiple identical copies of a read can be present as an artifact of PCR amplification procedure and this can provide false evidence for variant site discovery. Therefore, in the first step, we retained only a single copy of each read (Figure 2A). This filter can also reduce confounding effects of systematically bad sequencing cycles within a read. In the second step, if multiple reads map to the same genomic position, we randomly selected only one read from each of the categories U0, U1 and U2 (Figure 2A). Thus, there can be at most three reads that map to the same genomic position (Figure 3A). The application of the above two filters (named together as ‘Redundant Reads Filter’ in Figure 1) should reduce false-positive rate of SNV discovery. Since there can be only a small number of unique and non-redundant genomic reads at the exon edges, we generated a library containing exon junctions to detect potential SNVs in these genomic regions, which increased the power
of SNV detection at the exon edges. We found that about 6% of all significant SNVs are detected due to exon-junction reads. The nonredundant reads were analyzed by our point mutation analyzer. A very small probability of observing multiple overlapping but noncoincident short sequence reads agreeing at a given mismatched genomic location by random chance is taken as the evidence in favor of the presence of a genuine SNV at that location (Figure 2B and ‘Materials and methods’ section).

The number of reads that align uniquely to the genome and exon junctions is shown in Table 1. We obtained about 27 million uniquely mapped 30-bp sequence reads for each sample. The resulting mean coverage of exonic regions is \( \sim 11 \times \). Since gene expression varies dramatically, we examined the distribution of coverage for all exonic sequences (Figure 4A). Our data indicate that with 26 million uniquely mapped non-redundant short sequence reads, about 40% of exonic regions were covered \( \geq 5 \) times.

We performed sequencing cost analysis for SNV detection (see ‘Materials and methods’ section). We show that at the stringency we use to call SNV (\( P\)-value = 10\(^{-9} \)), fold coverage of \( C = 5 \) and \( C = 14 \) are needed to detect homozygous and heterozygous SNVs, respectively. At the sequencing depth we achieved (around 13 million 30-bp unique nonredundant reads), these fold coverages correspond to RPKM values 13 and 35, respectively. Thus, we estimate that about 40% of homozygous and 14% of heterozygous expressed SNVs were detected in this work. Our analysis demonstrates that about 80% of homozygous and 55% of heterozygous SNVs in expressed exons can be detected using 67 million 30-bp nonredundant unique reads (Figure 4B). However, our hypothesis is that mutation of a highly expressed gene may have more functional consequence than a gene expressed at low level or not expressed; therefore, it may not be necessary to do much deeper sequencing than what we have achieved in this study.

SNV validation and annotation

At a very stringent significance threshold (\( P\)-value < 1.0 \( \times 10^{-5} \)), we detected 12176 and 10621 SNV in Jurkat and CD4\(^{+} \) T cells, respectively. Many of detected sites overlap with known single nucleotide polymorphism sites (dbSNP build 126): 7473 for Jurkat and 7669 for CD4\(^{+} \) T cells (Figure 5A). Interestingly, more nonsynonymous SNVs in Jurkat cells as compared to CD4\(^{+} \) T cells (Figure 5B and Tables 1, S1 and S2 for further details), which could be related with the disease or generated during in vitro culture.

To validate the genetic mutations detected using RNA-Seq, we randomly selected five nonsynonymous SNVs that are also present in dbSNP and four SNVs that are novel in Jurkat cells (Table 2). The genomic regions containing these SNVs were amplified using PCR and sequenced using Sanger sequencing method. Our results indicate that all the nine SNVs were confirmed (Figure S1).

Figure 5. Summary of results. (A) Venn diagram of single nucleotide variants (SNVs) detected in Jurkat and CD4 samples. (B) Summary table of SNVs detected in Jurkat and CD4 samples. Shown in the brackets are numbers of SNVs that are novel, i.e. not present in dbSNP Build 126 database.

sequencing revealed that both the wild-type and mutated alleles were present, suggesting that only one parental copy is mutated and it is the mutated allele but not the wild-type allele that is expressed in Jurkat cells.

Among all the 12176 SNVs identified in Jurkat cells, 4703 are novel and 7473 are known (Figure 5B). Among these, we detected 3206 nonsynonymous and 47 nonsense mutations. Further analysis of the 47 nonsense SNVs indicates that 41 are novel. Interestingly, all the 20 Jurkat-specific nonsense SNVs are single-allele changes (Table 3). We were able to PCR amplify genomic regions containing 18 of these 20 SNVs and obtained their sequences using Sanger sequencing method. Our results indicate that 16 SNVs were confirmed (Figure S1).

Interestingly, we found that one of the two SNVs not confirmed by sequencing of genomic DNA was in fact present in mRNA as revealed by Sanger sequencing of cDNA (Figure S2). The SNV is located in the last exon of TAF6 gene. These results suggest that the SNV may be introduced by RNA-editing.

Mutations affecting splicing events are linked to a large number of genetic diseases (2). RNA-seq can be used to detect SNVs overlapping with potential splicing sites if the corresponding intron sequence is present at a sufficiently high level in cDNA. This can happen if (1) the gene is highly expressed so that there is a sufficient amount of pre-spliced RNA present, and/or (2) as a consequence of the splice-site mutation itself the intron is retained due to
aberrant splicing. We identified 66 and 31 SNVs in Jurkat and CD4 cells, respectively, which overlapped with splice sites and may lead to aberrant splicing process (Figure 5B). However, using exon body and junction reads, we did not observe any significant Jurkat-specific aberrant splicing events such as exon-skipping and intron retention at the sites of splicing variants (data not shown).

**CONCLUSIONS**

Even though whole-genome resequencing can provide unbiased information on SNVs in the human genome, it is still very expensive and time consuming. Additionally, the vast majority of genetic mutations linked to human diseases have been found in protein-coding regions, which make less than 2% of the human genome. Therefore, sequencing methods that are selective for protein-coding regions may significantly increase the efficiency and reduce the cost of identifying disease-related mutations. Since a genetic disease may be related with increased or decreased expression of a protein or a change in protein structure, sequencing the mRNAs may be a satisfying strategy because it can reveal the information of both gene expression and DNA sequence changes in the protein-coding regions of the human genome.

In this report, we have demonstrated that analyzing mRNAs using the next-generation sequencing techniques is an efficient and cost-effective method to identify SNVs in the protein-coding regions of the human genome, in addition to providing the expression information of the genome. Hence, it can be applied to identifying DNA changes in genetic diseases associated with expressed

### Table 2. Confirmation of selected Jurkat single nucleotide variants by Sanger sequencing of genomic DNA

| Gene   | Chromosome | Position* | Predicted alleleb | Reference allelec | #A | #C | #G | #T | P-value | Known SNP | Amino acid change | Confirmed |
|--------|------------|-----------|-------------------|-------------------|----|----|----|----|---------|-----------|------------------|----------|
| LCP1   | chr13      | 45606292  | C                  | T                 | 0  | 58 | 0  | 0  | 1.0e-102| Yes       | K → E           | Yes      |
| LOC554226 | chr2       | 132729041 | C                  | T                 | 2  | 53 | 1  | 1  | 1.9e-97 | No        | intronic         | Yes      |
| ECH1   | chr19      | 44013927  | G                  | T                 | 0  | 0  | 55 | 1  | 1.1e-95 | Yes       | E → A           | Yes      |
| SEPT9  | chr17      | 7306300   | G                  | A                 | 0  | 1  | 50 | 0  | 2.1e-90 | Yes       | M → V           | Yes      |
| POLR3K | chr16      | 43517     | C                  | A                 | 0  | 48 | 2  | 0  | 1.2e-88 | Yes       | S → A           | Yes      |
| CYC1   | chr8       | 14522820  | G                  | A                 | 0  | 0  | 49 | 0  | 7.0e-87 | Yes       | M → V           | Yes      |
| FLNA   | chrX       | 153235779 | A                  | G                 | 45 | 3  | 2  | 0  | 4.7e-82 | No        | R → W           | Yes      |
| MYO1G  | chr7       | 44983146  | T                  | C                 | 0  | 0  | 3  | 36 | 2.7e-69 | No        | Y → M           | Yes      |
| TAL1   | chr1       | 47456811  | T                  | C                 | 0  | 0  | 3  | 39 | 2.7e-69 | No        | UTR            | Yes      |

*Shows 1-based chromosomal location of SNV.

bShows the allele inferred from RNA-seq data using the Point Mutation Analyzer.

cShows the allele from hg18 (NCBI Build 36) human genome sequence; both alleles refer to the forward strand of the genome sequence.

#X' denotes the number of uniquely mapped nonredundant RNA-seq reads that have nucleotide X at the location of SNV.

*Known SNP* status is based on dbSNP build 126 database.

### Table 3. Jurkat-specific nonsense mutations

| Chromosome | Position | Gene | Mutant allele | WT allele | Jurkat reads variant:WT | CD4 reads variant:WT | Confirmed |
|------------|----------|------|---------------|-----------|-------------------------|----------------------|----------|
| chr22      | 37212044 | DDX17| A             | G         | 32:33                   | 0.60                 | Yes      |
| chr16      | 29995232 | PPP4C| T             | C         | 28:34                   | 0.22                 | Yes      |
| chr14      | 90841906 | CACD88C| A          | G         | 18:17                   | 1.37                 | Yes      |
| chr4       | 39921888 | RHOH | T             | C         | 19:21                   | 0.47                 | Yes      |
| chr17      | 77118842 | C17orf70| A         | G         | 17:19                   | 0.19                 | Yes      |
| chr19      | 1426114  | C19orf25| A         | G         | 16:11                   | 0.13                 | Yes      |
| chr20      | 61840255 | LIME1| T             | G         | 18:25                   | 0.05                 | Yes      |
| chr19      | 12647448 | MORG1| T             | C         | 15:12                   | 0.11                 | Yes      |
| chr14      | 77004596 | AHS1A| A             | G         | 14:49                   | 0.25                 | Yes      |
| chr1       | 15190223 | ILF2 | A             | G         | 17:50                   | 0.24                 | Yes      |
| chr8       | 144728307| NAPRT1| A            | G         | 13:14                   | 0.11                 | Yes      |
| chr20      | 62175757 | RGS19 | A             | G         | 12:15                   | 0.46                 | Yes      |
| chr9       | 130824412| SH3GLB2| A           | G         | 11:23                   | 0.18                 | Yes      |
| chr12      | 11910858 | GCN1L1| A            | G         | 10:16                   | 0.16                 | Yes      |
| chr17      | 77642429 | FASN | A             | G         | 11:40                   | 0.13                 | —        |
| chr12      | 8126211  | NECAP1| A             | C         | 8:9                     | 0.21                 | Yes      |
| chr7       | 99543029 | TAF6 | A             | G         | 9:24                    | 0.24                 | Yes      |
| chr17      | 40583130 | HEXIM1| T            | C         | 9:25                    | 0.42                 | Yes      |
| chr5       | 79534492 | SERINC5| T           | G         | 8:17                    | 0.17                 | No       |
| chr9       | 33437492 | AQP3 | A             | C         | 10:26                   | 0.65                 | —        |

*aConfirmed by Sanger sequencing of cDNA.*
aberrant protein products resulting from inherited or somatic point mutations.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
After the initial submission of the manuscript to NAR, a referee informed us that massive-scale mRNA sequencing had been used before for SNP detection in mouse stem cells (8) and in human HeLa cells (9).

FUNDING
The Intramural Research Program of the NIH, National Heart, Lung and Blood Institute. Funding for open access charge: The Intramural Research Program of the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES
1. Ley, T. J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 456, 66–72.
2. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeyesinghe, S., Krawczak, M. and Cooper, D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat., 21, 577–581.
3. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628.
4. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321, 956–960.
5. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 320, 1344–1349.
6. Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J. and Bähler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, 453, 1239–1243.
7. Baer, R. (1993) TAL1, TAL2 and LYL1: a family of basic helix-loop-helix proteins implicated in T cell acute leukaemia. Semin. Cancer Biol., 4, 341–347.
8. Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. et al. (2008) Stem cell transcriptome profiling via massively parallel RNA sequencing. Nat. Methods, 5, 613–619.
9. Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J.M. and Marra, M.A. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques, 45, 81–94.