

Accurate Estimation of Expression Levels of Homologous Genes in RNA-seq Experiments

BOGDAN PAȘANIUC,¹ NOAH ZAITLEN,¹ and ERAN HALPERIN^{2,3,4}

ABSTRACT

Next generation high-throughput sequencing (NGS) is poised to replace array-based technologies as the experiment of choice for measuring RNA expression levels. Several groups have demonstrated the power of this new approach (RNA-seq), making significant and novel contributions and simultaneously proposing methodologies for the analysis of RNA-seq data. In a typical experiment, millions of short sequences (reads) are sampled from RNA extracts and mapped back to a reference genome. The number of reads mapping to each gene is used as proxy for its corresponding RNA concentration. A significant challenge in analyzing RNA expression of homologous genes is the large fraction of the reads that map to multiple locations in the reference genome. Currently, these reads are either dropped from the analysis, or a naive algorithm is used to estimate their underlying distribution. In this work, we present a rigorous alternative for handling the reads generated in an RNA-seq experiment within a probabilistic model for RNA-seq data; we develop maximum likelihood-based methods for estimating the model parameters. In contrast to previous methods, our model takes into account the fact that the DNA of the sequenced individual is not a perfect copy of the reference sequence. We show with both simulated and real RNA-seq data that our new method improves the accuracy and power of RNA-seq experiments.

Key words: algorithms, gene searching, genetic mapping, genetic variation.

1. INTRODUCTION

NEXT GENERATION HIGH-THROUGHPUT SEQUENCING (NGS) technologies are rapidly establishing themselves as powerful tools for assaying a growing list of cellular properties including sequence and structural variation, RNA expression levels, alternative splice variants, protein-DNA/RNA interaction sites, and chromatin methylation state (Wang et al., 2009; Schuster, 2008; Marioni et al., 2008; Mortazavi et al., 2008; Johnson et al., 2007; Cokus et al., 2008). NGS enables thousands of megabases of DNA to be sequenced in a matter of days with very low cost compared to traditional Sanger sequencing.

¹Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts.

²International Computer Science Institute, Berkeley, California.

³Molecular Microbiology and Biotechnology Department and ⁴The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel.

It provides tens of millions of short reads, which can then be mapped back to a reference genome or used for de novo assembly. The advantages offered by NGS are underlined by the sheer wealth of significant novel discoveries not possible with existing chips and prohibitively expensive with previous sequencing methods.

As with any new technology, there are a host of new problems to solve in order to maximize the benefit of the data produced. In the case of NGS, many of the new methods adapt classic problems such as alignment and assembly to the relatively short, inaccurate, and abundant set of reads. Other methods, such as the one presented here, aim at optimizing the analysis of NGS assays previously done using microarray-based technologies such as quantifying gene expression levels from RNA data (*RNA-seq*). A first step in such an analysis is mapping the reads to a reference genome and aggregating the counts for each genomic location. Under the assumption that NGS samples short reads at random from the sequenced sample, the sequences with higher concentration will produce more reads. In the case of arrays, this corresponds to a higher probe intensity. Indeed, it was recently shown that the RNA-seq read counts and expression array probe intensities are highly correlated measurements for RNA expression levels (Mortazavi et al., 2008; Marioni et al., 2008).

Accurate estimation of the number of reads mapped to each genomic location critically depends on finding the location on the reference genome from which each read originated. While the majority of the reads produced by an NGS experiment map to a unique location along the genome, due to short read length, sequencing errors, and the presence of repetitive elements and homologs, a significant percentage of reads (up to 30% from the total mappable reads) are mapped to multiple locations (*multireads*). In the vast majority of RNA-seq experiments that have been published so far, the analysis consisted of simply disregarding the multireads from subsequent analyses. However, as previously noted (Mortazavi et al., 2008), if the multireads are discarded, the expression levels of genes with homologous sequences will be artificially deflated. If the multireads are split randomly amongst their possible loci, differences in estimates of expression levels for these genes between conditions will also be diminished leading to lower power to detect differential gene expression. Several groups have proposed a more intuitive alternative for dealing with multireads (Hashimoto et al., 2009; Mortazavi et al., 2008). Although there are small differences, they both adopt a heuristic approach, dividing the multireads amongst their mapped regions according to the distribution of the uniquely mapped reads in those regions. Intuitively, if there is a unique segment in the homologous region, then the distribution of the multireads in the repetitive segment of the region will follow the same distribution as the reads in the unique segment. This approach, although intuitive, is not optimal, as it does not thoroughly model the contribution of the multireads.

We note that a number of recent articles (Nicolae et al., 2010; Li et al., 2010; Jiang and Wong, 2009; Guttman et al., 2010; Trapnell et al., 2010) address related problems for the inference of expression levels using NGS data. The methods of Guttman et al. (2010) and Trapnell et al. (2010) address the inference of the transcripts using gapped alignments of reads across splice junctions aggregating reads into transcript structures followed by inference of expression levels of the inferred transcripts. Trapnell et al. (2010) use a Bayesian inference procedure based on importance sampling for estimating the abundance levels of the inferred transcripts. In this work, we assume the genome is fully annotated, and thus we do not infer the transcripts but focus only on estimation of their abundance levels. Particularly, we focus on solving the ambiguity in gene expression levels due to reads mapping to multiple locations in the genome. However, ambiguity can exist in the form of reads mapped to the same gene but coming from different isoforms. Several methods have been proposed for isoform expression inference (Nicolae et al., 2010; Li et al., 2010; Jiang and Wong, 2009; Trapnell et al., 2010) ranging from Poisson modeling of NGS data to Bayesian Network modeling of the short read data. All these methods assume no difference between reference genome and the genome in the experiment and treat all mismatches as sequencing errors. Our work focuses specifically on resolving the ambiguity of homologous gene expression levels due to reads mapped to multiple locations, allowing for variation in the genomic data as opposed to the reference genome.

In this work, we propose a rigorous framework for handling multireads that is applicable to several different assays including RNA-seq. In contrast to previous approaches, which were heuristic in their nature, we propose a generative model that describes the results of an RNA-seq experiment including multireads. An important feature of our model is that it takes into account genetic variation between the reference human genome sequence and the sequence of the studied sample, improving accuracy in

some instances and allowing for simultaneous expression analysis and genotyping. We further developed algorithms for estimating the parameters of the model using a maximum likelihood approach. We show through simulations and real RNA-seq data that our method significantly improves the accuracy and power of detecting differentially expressed genes under several measures. Particularly, our results on real data demonstrate that, in an RNA-seq experiment comparing two tissues, we can potentially discover many more genes that are differently expressed between the tissues. In addition, our treatment of genetic variation allows us to simultaneously call variants (e.g., locations where the sequenced sample varies from reference), and use the location of these variants to further resolve the location of the multireads.

An implementation of our method is freely available for download as part of the software package SeqEm at <http://seqem.icsi.berkeley.edu/seqem/>.

2. METHODS

We will first describe our probabilistic generative model for an RNA-seq experiment. Let $G = (G_1, \dots, G_n)$ be n contiguous DNA regions representing genes or other potentially expressed sequences. For each G_i , we define the RNA cellular concentration of the gene as P_i , s.t. $\sum_{i=1}^n P_i = 1$. $P = (P_1, \dots, P_n)$ can be interpreted as the normalized expression levels for the regions in G . Our model assumes that reads of length l are generated by randomly picking a region R from G according to the distribution P , and then copying l consecutive positions from R starting at a random position in the gene. The copying process is error-prone, with probability $\varepsilon(k)$ for a sequencing error in the k^{th} position of the read. The model is easily adapted to multi-length reads, but a fixed length is used here for simplicity. This process is repeated until we have a set of m reads $R = r_1, \dots, r_m$ generated according to the model described above. The objective of an RNA-seq experiment is to infer P from R .

The first step in an RNA-seq experiment consists of mapping the results of an NGS run to the reference genome. Mapping methods such as ELAND (<http://www.illumina.com>), Maq (Li et al., 2008), and bwa (Li and Durbin, 2009) provide for each read its most probable alignment, its position, and how many mismatches the alignment contains. Due to sequencing errors, some reads may not align perfectly. Furthermore, multireads align to more than one position, especially if the sequenced regions overlap with repeated genomic sequences such as homologous genes or repeats like ALUs, LINES, and SINES.

In the context of our model, each read r_i originated from one of the regions in G , but due to sequencing errors it may not align perfectly to that region; furthermore, due to repeated sequences, it may also align to other regions. Put differently, for each region G_j and read r_i , we have a probability $p_{ij} = P(r_j|G_i)$, the probability of observing r_j given that the locus of the read was gene G_i . In practice, for each read r_j , this probability will be close to zero for all but a few regions. The likelihood of observing the m reads can be written as:

$$L(P; R) = \prod_{j=1}^m P(r_j|G, P) = \prod_{j=1}^m \sum_{i=1}^n P(G_i)P(r_j|G_i) = \prod_{j=1}^m \sum_{i=1}^n P_i p_{ij}$$

Unfortunately, we do not know the expression levels P . A natural way of finding estimates for P is given in the following problem formulation for the Maximum Likelihood Expression Inference (MLEI) problem:

Definition 1 (MLEI). *Given a set of reads r_1, \dots, r_m and a set of regions G_1, \dots, G_n , find a probability P_i for every region G_i so that $\sum_i P_i = 1$, and so that the likelihood of the data $L = \prod_{j=1}^m \sum_{i=1}^n P_i p_{ij}$ is maximized.*

As shown in Halperin and Hazan (2006), the likelihood objective function is concave, and the maximization of this function is polynomially solvable since there is a separation oracle as long as the p_{ij} coefficients are fixed. We present here an Expectation-Maximization (EM) algorithm for the MLEI problem. Since this problem is concave, the EM algorithm will converge to the optimal solution.

2.1. EM algorithm for inferring expression levels

We now describe an algorithm for solving the MLEI problem. We are searching for $P = \{P_1, P_2, \dots, P_n\}$ such that the likelihood of the data is maximized. Let M be the underlying true unobserved matching of

reads to regions. Then the following is an EM algorithm that searches for P that maximizes $L(P; R)$. Let $P^{(t)}$ be the current estimate of P .

E step:

$$\begin{aligned} Q(P|P^{(t)}) &= E_{M|R, P^{(t)}}[\log L(P; R, M)] \\ &= E_{M|R, P^{(t)}}\left[\sum_{i=1}^m (\log P_{M(i)} + \log p_{iM(i)})\right] \\ &= \sum_{i=1}^m \sum_{j=1}^n \left[(\log P_j + \log p_{ij}) \times \frac{P_j^{(t)} p_{ij}}{\sum_{j=1}^n P_j^{(t)} p_{ij}} \right] \end{aligned}$$

M step:

$$\begin{aligned} P^{(t+1)} &= \arg \max_P Q(P|P^{(t)}) \\ &= \arg \max_P \left[\sum_{i=1}^m \sum_{j=1}^n a_{ij} \log P_j + \sum_{i=1}^m \sum_{j=1}^n a_{ij} \log p_{ij} \right] \end{aligned}$$

where $a_{ij} = \frac{P_j^{(t)} p_{ij}}{\sum_{j=1}^n P_j^{(t)} p_{ij}}$. Given that p_{ij} (the probability of read j if it came from region j) are fixed, maximizing the above function reduces to finding

$$P^{(t+1)} = \arg \max_P \sum_{i=1}^m \sum_{j=1}^n a_{ij} \log P_j = \sum_{j=1}^n \left(\sum_{i=1}^m a_{ij} \right) \log P_j$$

It can be easily shown that the maximum is achieved at:

$$P_j^{(t+1)} = \frac{\sum_{i=1}^m a_{ij}}{\sum_{i=1}^m \sum_{j=1}^n a_{ij}}, \forall j$$

Since the likelihood function is concave (Halperin and Hazan, 2006), the above EM is guaranteed to converge to the optimal solution. Although it does not have the same polynomial time guarantee as the method in Halperin and Hazan (2006), in practice it outperforms their HAPLOFREQ method. It also provides a basic framework for the extension of the MLEI problem to the case of joint estimation of expression levels and variants where the sequenced sample differs from the reference genome. Since single-nucleotide polymorphisms (SNPs) are the most common source of variation in the human genome, we focus primarily on single nucleotide variants although other type of variants can be easily incorporated into the model. The model of reads with SNP variants is more realistic and may also be more powerful for certain cases, since SNPs can be used to distinguish genomic locations in homologous regions. We demonstrate in the Results section that the solution obtained by the EM estimates the gene expression levels P more accurately than the heuristic methods of either ignoring the multireads altogether or dividing them among the regions they map to.

2.2. Joint estimation of expression levels and SNP variants

In the above formulation, we implicitly assumed that the probabilities p_{ij} were fixed and easy to compute since we had a fixed reference dataset. All differences between reads and reference were assumed to be due to errors, and p_{ij} was simply a function of our model parameters. In practice, however, the sequenced DNA may be slightly different than the reference genome, particularly in SNP positions. To model the SNP locations, we introduce a variable $X_k = \{X_k^1, X_k^2\}$ with $X_k^1, X_k^2 \in \{A, C, T, G\}$ for each genomic position k , which denotes the genotype of the sequenced sample at that location. The values of X_k are unknown, and they have to be inferred. We can assume that we have a prior distribution of X_k that corresponds to the distribution of the allele frequencies in the genome; this distribution can be empirically estimated (depending on the ancestry of the sample) from the HapMap (The International HapMap Consortium, 2007) data and particularly the ENCODE (The ENCODE

Project Consortium, 2007) regions, as well as the 1000 genomes project when the data becomes available. Particularly, we can have an estimate of the distribution of allele frequency across positions that are not known to be SNPs based on the ENCODE regions, and for the other positions we have their allele frequencies from dbSNP or from HapMap. Now, if the plausible alignment of read r_i to region G_j spans the positions X_1, \dots, X_l , assuming that sequencing errors are independent of each position, we can write p_{ij} as:

$$p_{ij} = \prod_k \gamma(X_k, r_i^k, k)$$

where

$$\gamma(X_k, r_i^k, k) = \begin{cases} \epsilon(k), & \text{if } X_k^1 \neq r_i^k, X_k^2 \neq r_i^k \\ 1 - \epsilon(k), & \text{if } X_k^1 = r_i^k, X_k^2 = r_i^k \\ 0.5, & \text{otherwise} \end{cases}$$

$\epsilon(k)$ is the error rate function in a read at position k . The dependency of the error rate on the position comes from technological constraints as the error rate is expected to increase with the length of the reads (see Dohm et al. [2008] for empirical estimates of Solexa error rates). Based on this, the problem of joint estimation of expression levels and SNP variants can be defined as follows:

Definition 2 (MLEI-SNP). Given a set of reads r_1, \dots, r_m and a set of regions G_1, \dots, G_n , find a probability P_i for every region G_i and genotype $X_k = \{X_k^1, X_k^2\} \in \{A, C, T, G\}^2$ for every location k , so that $\sum_i P_i = 1$, and so that the likelihood of the data $L = \prod_{j=1}^m \sum_{i=1}^n P_i p_{ij}$ is maximized, where $p_{ij} = \prod_{k=1}^l \gamma(X_k, r_i^k, k)$.

EM extension with SNP variants. In order to maximize the likelihood of the data, we are now looking for both $P = \{P_1, P_2, \dots, P_n\}$ s.t $\sum P_i = 1$ and genotype calls $X = \{x_1, \dots, x_k\}$ for every genomic location so that the likelihood of the data $L(P, X; R) = \prod_{j=1}^m \sum_{i=1}^n P_i p_{ij}$ is maximized, where p_{ij} is defined as before:

$$p_{ij} = \prod_k \gamma(X_k, r_i^k, k)$$

The EM algorithm can be adapted as follows:

E step:

$$\begin{aligned} Q(P, X | P^{(t)}, X^{(t)}) &= E_{M|R, P^{(t)}, X^{(t)}} [\log L(P, X; R, M)] \\ &= E_{M|R, P^{(t)}, X^{(t)}} \left[\sum_{i=1}^m \log P_{M(i)} p_{iM(i)} \right] \\ &= \sum_{i=1}^m \sum_{j=1}^n \left[(\log P_j p_{ij}) \times \frac{P_j^{(t)} P_{ij}^{X^{(t)}}}{\sum_{j=1}^n P_j^{(t)} P_{ij}^{X^{(t)}}} \right] \end{aligned}$$

M step:

$$\begin{aligned} (P^{(t+1)}, X^{(t+1)}) &= \arg \max_{P, X} Q(P, X | P^{(t)}, X^{(t)}) \\ &= \arg \max_{P, X} \left[\sum_{i=1}^m \sum_{j=1}^n a_{ij} \log P_j p_{ij} \right] \\ &= \arg \max_{P, X} \left[\sum_{i=1}^m \sum_{j=1}^n a_{ij} \log P_j + \sum_{i=1}^m \sum_{j=1}^n a_{ij} \log p_{ij} \right] \end{aligned}$$

Since the two terms in the above equation are independent, we can maximize them separately. Just as before the first term in the equation above is maximized when $P_j^{(t+1)} = \frac{\sum_{i=1}^m a_{ij}}{\sum_{i=1}^m \sum_{j=1}^n a_{ij}}$, where $a_{ij} = \frac{P_j^{(t)} P_{ij}^{X^{(t)}}}{\sum_{j=1}^n P_j^{(t)} P_{ij}^{X^{(t)}}}$.

The second term is more complicated as we need to find X^* that maximizes $\sum_{i=1}^m \sum_{j=1}^n a_{ij} \log p_{ij}$. However, since the term depending on p_{ij} is a log of a product, we can decompose it into independent contributions for each genomic location k and optimize each X_k independently. Namely,

$$\begin{aligned} \sum_{i=1}^m \sum_{j=1}^n a_{ij} \log p_{ij} &= \sum_{i=1}^m \sum_{j=1}^n a_{ij} \log \prod_k \gamma(X_k, r_i^k, k) \\ &= \sum_{i=1}^m \sum_{j=1}^n a_{ij} \sum_k \log \gamma(X_k, r_i^k, k) \\ &= \sum_k \sum_{\text{read } i \text{ spans } k} a_{ij} \log \gamma(X_k, r_i^k, k) \end{aligned}$$

and thus we set

$$X_k^{(t+1)} = \arg \max_{X_k = (x_k^1, x_k^2)} \sum_{\text{read } i \text{ spans } k} a_{ij} \log \gamma(X_k, r_i^k, k)$$

In practice, we can speed up the computations by noticing that, in the M step when finding new estimates for X_k^{t+1} , we only need to consider locations k , at which there are at least $c > 0$ mismatches to the reference.

3. RESULTS

In this section, we present results on both simulated and real data sets showing the superior accuracy of our approach when compared to three previously proposed heuristic approaches for this problem. The first method we compare to is the standard method that ignores all multireads and estimates the expression levels P_i^{uniq} as the percentage of unique reads mapped to region i amongst all uniquely mapped reads. The second method estimates P_i by dividing the ambiguous reads uniformly between each region it maps to. Namely, $P_i^{unif} = \frac{1}{m} \sum_{j: j \text{ maps to } i} \frac{1}{h(i)}$, where $h(i)$ is the number of locations read r_i maps to. A more intuitive approach (Hashimoto et al., 2009; Mortazavi et al., 2008) is to divide each read amongst each location it maps to according to weights, where the weights are given by the distribution of the uniquely mapped reads in those regions; we denote this method as the *weighted* approach.

Performance measures. We use two correlated measures for the distance between the estimated and true distributions of the RNA expression levels P . P_i denotes the true expression level of a gene and \hat{P}_i is the estimated expression level. The first measure we use, the *error rate*, is computed as $\frac{1}{n} \sum_i \frac{|P_i - \hat{P}_i|}{P_i}$, and it quantifies the average distance between the true and the estimated expression level in a region. A second approach to measure the accuracy of the estimates is the “goodness-of-fit” measure between the two distributions, in terms of *chi-square difference*: $\sum_i \frac{(P_i - \hat{P}_i)^2}{P_i}$. This measure is of particular interest as it is correlated to the power to detect differentially expressed regions.

Simulated datasets. In the first set of experiments, we assessed the performance of our framework on RNA-seq by simulating short reads based on chromosome 1 from the human genome as a reference sequence. We focused on known homologous genes, since they are the genes that are most affected by multireads. To do this, we downloaded the 756 human homologous genes from chromosome 1 from the Homologene (<http://www.ncbi.nlm.nih.gov/homologene/>) database. We removed all overlapping genes and genes with no other homologs in human resulting in 51 genes over 95kb.

The human reference genome does not contain information about possible polymorphisms, however it is expected that we will see both homozygous and heterozygous variants when sequencing a random individual in comparison to the reference. Given that the sequencing sample is different from the reference at a locus where the SNP allele frequency is f , the probability for a heterozygote is $2f(1-f)$ and for a homozygous variant different from the reference is $f^2(1-f) + f(1-f)^2 = f(1-f)$. Thus, given that a site is different from the reference, the probability of a heterozygote is $2/3$, and of a homozygote is $1/3$, regardless of the allele frequency f . As done elsewhere (Li et al., 2008), we used this observation when simulating a sample. First we pick a set of variants (where the sample differs from reference) with a rate of 10^{-3} (which is the approximate frequency of SNPs in the genome) and then we randomly set $2/3$ of the variants as heterozygous and $1/3$ homozygous. In order to make the simulations as close to the actual data as possible, we also picked genotypes for the sample at known HapMap SNPs from the distribution given by the HapMap CEU frequencies.

For each of the 51 homologous genes, we randomly chose P_i according to the uniform distribution, and normalized so that $\sum_i P_i = 1$; P_i represents the true expression rate for gene i . We generated x_i reads for this region, where $x_i = \frac{C \times L(i) \times P_i}{T}$. C is a parameter of the simulation denoting the coverage rate, $L(i)$ is the length of the gene in base-pairs (we only count the exons) and T is the length of the read. Although currently available NGS technologies such as Solexa (<http://www.illumina.com>) or ABI SOLID (<http://solid.appliedbiosystems.com/>) produce reads of length 20–40 base-pairs it is expected that the read length will increase dramatically to up to 100bp and more in the near future. For this reason, we use simulations for two tag lengths ($T=32$ and $T=100$) thus simulating both currently available technologies and future technological developments. For each read at every location we inserted errors using a rate of $\varepsilon = 0.01$; similar results were obtained on simulations using an empirical error model that was estimated by Dohm et al. (2008) (data not shown). The reads were mapped to chromosome 1 hg18 using the bwa (Li and Durbin, 2009) mapping algorithm with default parameters.

Inferring expression levels in homologous genes. In our first set of results, we compared the EM algorithms with or without SNP variant calling to previously employed methods. Figure 1 shows that both EM algorithms outperform the other methods for both 32- and 100-bp length reads as well as for the different accuracy measures. Indeed for reads of length 32 the error rate decreases from approximately 30% for the *uniq* method that uses only the uniquely mapped reads to approximately 20% for both EM methods. The improvement, although still substantial, is more modest for reads of length 100, probably due to a smaller number of multireads as compared to reads of length 32.

To further highlight the effect of including the multireads in subsequent analyses as opposed to the general approach of using only the uniquely mapped reads, we assessed the quality of SNP variant inference with or without multireads. To maintain a meaningful comparison, we called SNP variants based on unique reads under the same likelihood method for calling SNPs as in the EM algorithm of Section 2.2. Table 1 shows the true and false positive rates for SNP variant calling showing that the *e-m-snps* method outperforms the *uniq* method for all studied coverages when compared to the method that employs only the uniquely mapped reads.

Detecting differential expression. Using the same set of genes as before we simulated pairs of experiments with different expression levels for the genes. Using the true expression levels and a standard chi-square test ($\alpha = 0.01$), we first computed a set of differentially expressed genes between the experiments

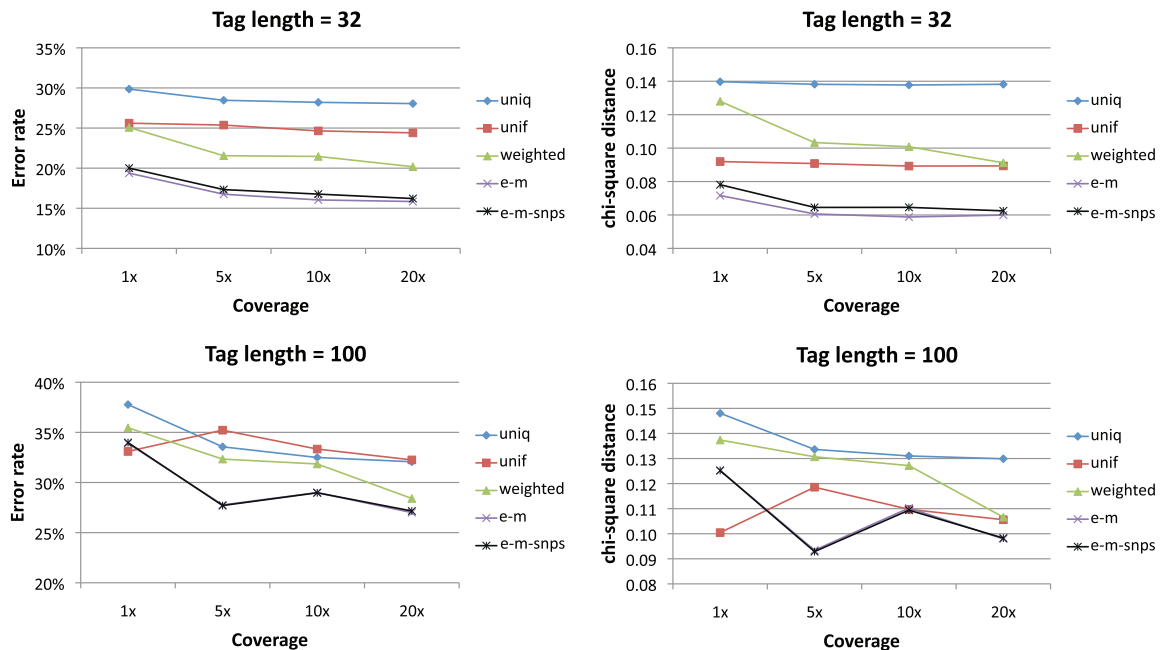


FIG. 1. Accuracy of gene expression inference based on simulated RNA-seq data for different read lengths and different accuracy measures. Results are given as averages over 100 simulated datasets. The EM methods outperform the heuristic methods of assigning reads as well as the approach of ignoring multireads.

TABLE 1. VARIANT CALLING RATES ON SIMULATED DATASETS WITH READS OF LENGTH 32 FOR VARIOUS COVERAGES

Coverage	Method	TPR	FPR
1×	uniq	18.00%	2.39E-05
	e-m-snps	18.26%	4.97E-05
5×	uniq	53.19%	3.27E-05
	e-m-snps	55.52%	3.99E-05
10×	uniq	69.67%	4.82E-05
	e-m-snps	73.55%	4.13E-05
20×	uniq	79.23%	3.50E-05
	e-m-snps	83.65%	2.26E-05

Results given in averages over 100 simulated datasets.

which serve as the gold standard “true” differentially expressed genes. We assessed the capacity of identifying the differentially expressed genes when different methods were used for estimating P_i^t s. The EM method shows the overall best performance, area under ROC curve of 0.83, compared to 0.75 for the *uniq* method and 0.81, 0.82 for the *unif* and *weighted* methods. For $\alpha = 0.05$ cutoff, EM achieves (true positive, false positive) rates of (97.5%, 24.5%)—compared to (88.4%, 20.8%) for *uniq* method, (95.9%, 26.6%) for *unif* method, and (96.6%, 26.4%) for *weighted* method.

Real dataset. We also applied our methods to a real RNA-seq data set from Marioni et al. (2008) consisting of two runs of an Illumina Genome Analyzer with half of the lanes containing human liver RNA and half kidney. We mapped all the reads with bwa (Li and Durbin, 2009) to the human genome sequence build hg18 and counted the number of reads in exons (we used the exon annotation of UCSC genome browser, <http://genome.ucsc.edu/>). The read counts per gene were highly correlated across lanes and did not exhibit a lane effect for most lanes (Marioni et al., 2008). We used the data from lanes one and two from the first run to estimate kidney and liver expression levels. We used the *weighted* method and our EM method to estimate the read counts for each gene. In this case, we do not know the true expression levels of the genes so we can not report which method is more accurate. Instead, we measure the number of genes exceeding a $5 \times \log_2$ fold change between each of the methods. The $5 \log_2$ fold change threshold we chose has the property that all genes exceeding this threshold in both the *weighted* and EM methods also exceed this threshold on the Affymetrix arrays. This suggests it is so conservative that it is 100% specific with 0 false positives. It would be useful to examine specificity and sensitivity at other thresholds but the true set of differentially expressed genes (i.e., a gold standard) for making such a comparison does not exist yet, and thus we restrict ourselves to the extreme $5 \log_2$ fold change.

For genes with uniquely mapped reads, these methods will perform identically, so we restricted our analysis to the 2207 genes with more than 200 multireads. For this set of homologous genes, our EM method found 94 highly differentially expressed genes, while the *weighted* method reported only 86, a decrease of 8.5%. All of the genes found to be highly differentially expressed using the *weighted* method were contained in the set found using EM. To verify that the additional eight genes we found using EM were not false positives, we examined their expression levels in the GeneAtlas project (Su et al., 2004), a comprehensive survey of gene expression in human tissues. For seven of the eight additional genes, we found that GeneAtlas expression levels were consistent with the EM findings; the probe intensities were greater than 50 in one tissue and less than 10 in the other. Figure 2 shows an example for the gene ENSG00000138075 (ABCG5). Note that ABCG5 has a known homolog ABCG8, so it is one of the cases that our method addresses. Only one of these eight genes predicted to be differentially expressed by EM was not differentially expressed in the GeneAtlas. Overall, these data confirm the increased power of our method, suggesting that the additional differentially expressed genes found by the EM are true positives.

4. CONCLUSION

Given the dropping cost of sequencing and the numerous advantages that RNA-seq has over expression array-based experiments, it is likely that in the near future RNA-seq will become a pervasive choice for measuring cellular RNA expression levels. Many of the analyses conducted so far have utilized varying

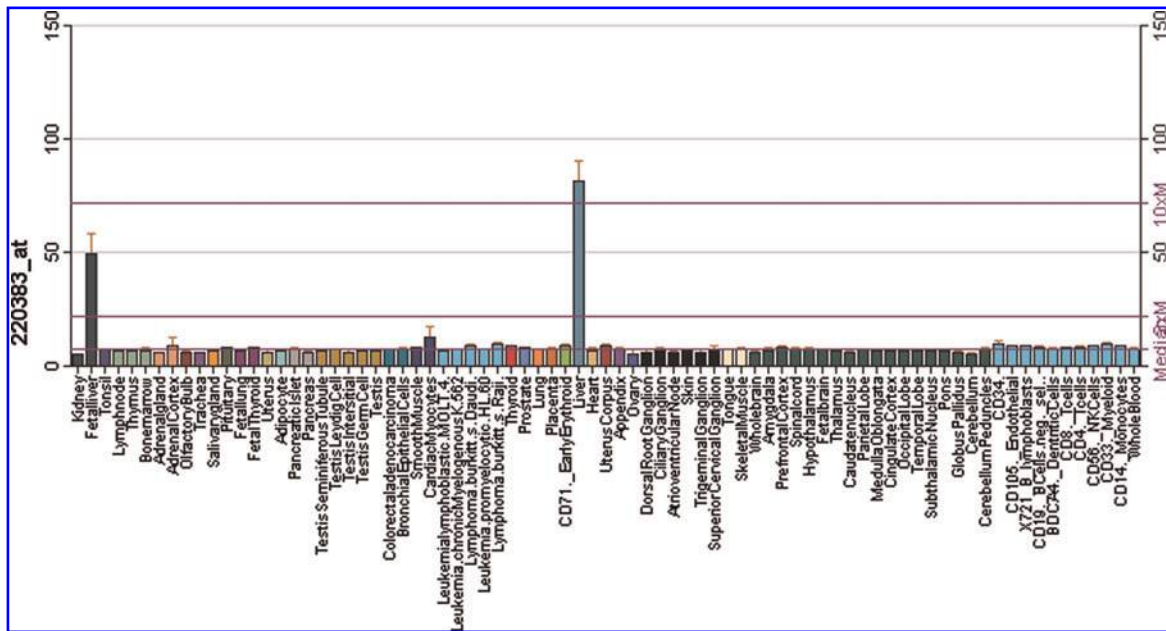


FIG. 2. Expression levels of gene ABCG5 in the GeneAtlas (<http://biogps.gnf.org>) project with high expression in liver and fetal-liver. Gene ABCG5 is shown to be highly differentially expressed between liver and kidney in Marion et al. (2008) RNA-seq data only when using our EM method for inferring gene expression levels.

methods, and it is currently unclear which strategies will prove to be the most accurate and powerful. Considering the rich literature discussing proper analysis of microarray data over the last fifteen years, it is likely that methods for this new technology can be significantly improved.

This work addresses an important aspect of RNA-seq analysis: how to handle reads from homologous and repetitive elements that map to multiple genomic locations. Our results clearly show that naive approaches significantly underestimate the true expression of homologous genes. Unlike previous heuristic approaches, we present methods based on a rigorous probabilistic generative framework for an RNA-seq experiment and show that our approach consistently outperforms all previous attempts at solving this problem. We also applied our approach on a real RNA-seq data set to find several new highly differentially expressed genes when compared to previous approaches; these findings were confirmed by existing expression array data sets.

We have identified several areas of improvement that we plan to address in future work. Currently, our method is limited to the use of consensus genes and may be improved by additionally modelling isoforms, splice variants, allelic heterogeneity, and un-annotated genes. In addition, the problem of multireads extends beyond RNA-seq experiments. For example, in both ChIP-seq and RIP-seq scenarios, array-based methods are replaced with an NGS approach, and so analysis methods must again handle multireads. Instead of determining the distribution of multireads as in RNA-seq, a binary signal is returned specifying whether or not a particular transcription factor binds to a specific genomic location. Solving the multiread problem in this context can potentially increase the power of detecting interesting loci, particularly when these loci fall within repetitive elements of the genome.

ACKNOWLEDGMENTS

E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. E.H. and B.P. were supported by National Science Foundation grant HS-071325412. E.H. and N.Z. were supported by the Israel Science Foundation grant no. 04514831.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Cokus, S.J., Feng, S., Zhang, X., et al. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215–219.
- Dohm, J.C., Lottaz, C., Borodina, T., et al. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105.
- Guttman, M., Garber, M., Levin, J.Z., et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 28, 503–510.
- Halperin, E., and Hazan, E. 2006. HAPLOFREQ: estimating haplotype frequencies efficiently. *J. Comput. Biol.* 13, 481–500.
- Hashimoto, T., de Hoon, M.J., Grimmond, S.M., et al. 2009. Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics* 25, 2613–2614.
- Jiang, H., and Wong, W.H. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25, 1026–1032.
- Johnson, D.S., Mortazavi, A., Myers, R.M., et al. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502.
- Li, B., Ruotti, V., Stewart, R.M., et al. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26, 493–500.
- Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858.
- Marioni, J.C., Mason, C.E., Mane, S.M., et al. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517.
- Mortazavi, A., Williams, B.A., McCue, K., et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628.
- Nicolae, M., Mangul, S., Mandoiu, I., et al. 2010. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Proc. WABI* 2687.
- Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18.
- Su, A.I., Wiltshire, T., Batalov, S., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101, 6062–6067.
- The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
- The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861.
- Trapnell, C., Williams, B.A., Pertea, G., et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515.
- Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Address correspondence to:
Dr. Bogdan Paşaniuc
Department of Epidemiology
Harvard School of Public Health
655 Huntington Avenue
Boston, MA 02115

E-mail: bogdan.pasaniuc@gmail.com