Rozov et al. BMC Bioinformatics 2012, 13(Suppl 6):S2
http://www.biomedcentral.com/1471-2105/13/S6/S2

BMC Bioinformatics

PROCEEDINGS Open Access

MGMR: leveraging RNA-Seq population data to optimize expression estimation

Roye Rozov[1], Eran Halperin[1,2,3*], Ron Shamir[1]

From Second Annual RECOMB Satellite Workshop on Massively Parallel Sequencing
Barcelona, Spain. 19-20 April 2012

Abstract

Background: RNA-Seq is a technique that uses Next Generation Sequencing to identify transcripts and estimate transcription levels. When applying this technique for quantification, one must contend with reads that align to multiple positions in the genome (multireads). Previous efforts to resolve multireads have shown that RNA-Seq expression estimation can be improved using probabilistic allocation of reads to genes. These methods use a probabilistic generative model for data generation and resolve ambiguity using likelihood-based approaches. In many instances, RNA-seq experiments are performed in the context of a population. The generative models of current methods do not take into account such population information, and it is an open question whether this information can improve quantification of the individual samples

Results: In order to explore the contribution of population level information in RNA-seq quantification, we apply a hierarchical probabilistic generative model, which assumes that expression levels of different individuals are sampled from a Dirichlet distribution with parameters specific to the population, and reads are sampled from the distribution of expression levels. We introduce an optimization procedure for the estimation of the model parameters, and use HapMap data and simulated data to demonstrate that the model yields a significant improvement in the accuracy of expression levels of paralogous genes.

Conclusions: We provide a proof of principal of the benefit of drawing on population commonalities to estimate expression. The results of our experiments demonstrate this approach can be beneficial, primarily for estimation at the gene level.

Introduction

With the rapid decline in the cost of sequencing, RNA-Seq has emerged as a legitimate competitor to mi-croarrays as a means of assessing global gene expression. Even as arrays currently enjoy a cost advantage, many new applications of information accessible only through sequencing further strengthen the case that sequencing may soon supplant arrays as the technology of choice for transcription analysis. One such application is fine-grained assessment of variation in expression and the sources for such variation, as exemplified by recent large-scale RNA-Seq studies [1,2] of two different HapMap [3] populations. Such studies complement genomic DNA sequencing by elucidating the link between SNPs and expression.

Unfortunately, with any new technology come its limitations, and several studies have pointed out that RNA-Seq is not immune to bias [4-6]. Perhaps the most widely discussed hurdle to accurate estimation in the case of RNA-Seq is the problem of reads mapped to multiple locations in the target genome (or in the target transcript sequences). These reads, which are called *multireads*, can stem from either paralogous gene sequences or from different isoforms of the same gene that share exons.

Several methods have emerged to address the multi-read problem for paralog and isoform estimation [7-10]. These methods are all based on probabilistic modeling

* Correspondence: heran@icsi.berkeley.edu
[1]The Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel
Full list of author information is available at the end of the article

that is optimized by an expectation maximization procedure. It has been repeatedly shown that by using such methods one can get better quantification of the expression levels compared to quantification based on naive approaches of read assignment.

In many applications, a set of samples is studied. For instance, one may be interested in comparing the expression levels in cases verses controls, or in tissues originating from different organs. In such cases, it is plausible that the commonality of expression patterns within each of the defined groups of studied samples may be used to improve the quantification results in each of the samples. We demonstrate that by analyzing expression profiles of a population together, one gets expression estimates more accurate than those obtained by estimating individual sample expression levels independently. We describe and implement a probabilistic model of the sequencing process that incorporates population commonalities, and demonstrate its advantages over existing methods in the population setting.

Methods

RNA-Seq multiread expression estimation

As we seek to extend the prevalent generative model of RNA-Seq [7-11], we begin by reviewing the basic elements of that model. Let $G = (G_1, ...,G_M)$ be the set of M transcribed regions considered and $P = (P_1, ...,P_M)$ be the proportions of RNA bases attributed to each transcript out of the total number of transcribed bases in a sequenced sample. Regions may be either genes or transcripts, depending on the level of transcription being investigated. We require P to satisfy $\sum_{g\epsilon G} P_g = 1$ and $\forall g\epsilon G, 0 \le P_g \le 1$.

The model describes an RNA sequencing experiment where regions in G are randomly chosen according to the distribution P, start positions in these regions are chosen uniformly, and reads of length ℓ are generated by copying ℓ consecutive bases from each chosen region to produce a set of reads $R = (r_1,..., r_\rho)$. Sequencing is assumed to be error prone, leading to a certain probability of error for each read base. Based on the repetitions present in the set of regions and errors in alignment, reads may fail to map to the region from which they originate or may map to additional locations. Thus, we assign a probability of obtaining read r_j given that it originated from region G_k, $P(r_j|G_k) \equiv \frac{(1-\varepsilon)^{(\ell - error_{jk})}\varepsilon^{error_{jk}}}{\ell_k}$. In this case we rely on the alignment of r_j to G_k to afford us the best match position instead of summing over all possible starting positions. ℓ_k is the effective length of G_k (i.e., the number of start positions from which a full length read can be derived) as defined in [11], ϵ is taken to be a constant per-base error rate, errors are assumed to be independent, and $error_{jk}$ is the number of mismatches in the best alignment of r_j to G_k.

This formulation leads to the likelihood of observing the data:

$$L(P;R) = \prod_{j=1}^{\rho} P(r_j|G,P) = \prod_{j=1}^{\rho}\sum_{k=1}^{M} P(G_k)P(r_j|G_k) \quad (1)$$

This likelihood function is used to estimate P given the read alignments. Typically, one will use expectation maximization to find the P for which the likelihood is maximized. It is assumed that $P(r_j|G_k)$ is zero for all regions to which r_j is not aligned.

Common population extension

To estimate expression levels in N individuals from a defined population, we modify the above model by assuming that samples are drawn from a common population. This is imposed by having $P = [(P_{11}, ...,P_{1M}), .., (P_{N1} ...,P_{N\ M})]$ be probability densities drawn from a common Dirichlet distribution, defined by a set of hyper-parameters specific to the population: $\forall i\epsilon[1, N]$, $\mathbf{p_i} = (P_{i1},..., P_{iM}) \sim Dirichlet\ (\alpha_1,..., \alpha_M)$.

For sample i, we denote the set of reads as $R_i = (r_{i1}, \ldots, r_{i\rho_i})$, where each r_{ij} is mapped to one or more regions in G. The output of a read alignment program defines the set of accepted regions for the read (in practice only alignments with up to 2 errors are accepted) and provides the number of errors in alignment for each read-region pair. This allows us to calculate $P(r_{ij}|G_k)$ as done above for one sample. For convenience we denote $P(r_{ij}|G_k) = q_{ijk}$ (taken to be zero for all regions not mapped to), which is independent of $\boldsymbol{\alpha}$ and $\boldsymbol{P}$.

As before, our objective is to estimate P, but in this case we must optimize by estimating P and $\boldsymbol{\alpha}$ together. We begin by writing the likelihood function:

$$L(\mathbf{p_1}, \ldots, \mathbf{p_N}, \boldsymbol{\alpha}; R) = Pr(\mathbf{p_1}, \ldots, \mathbf{p_N}|\boldsymbol{\alpha})Pr(\boldsymbol{R}|\mathbf{p_1}, \ldots, \mathbf{p_N}) \quad (2)$$

Since expression values are sampled from the Dirichlet distribution,

$$Pr(\mathbf{p_1}, \ldots, \mathbf{p_N}|\alpha) = \prod_{i=1}^{N} P(\mathbf{p}_i|\alpha) = \prod_{i=1}^{N} C(\alpha)\prod_{k=1}^{M} P_{ik}^{\alpha_k - 1} \quad (3)$$

Where

$$C(\alpha) \equiv \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \quad (4)$$

and similar to (1) above,

$$Pr(\boldsymbol{R}|\mathbf{p_1}, \ldots, \mathbf{p_N}) = \prod_{i=1}^{N}\prod_{j=1}^{\rho_i}\sum_{k=1}^{M} P_{ik}q_{ijk} \quad (5)$$

This leads to

$$L(\mathbf{p}_1,\ldots,\mathbf{p}_N,\alpha;R)=\left[\prod_{i=1}^{N}C(\alpha)\prod_{k=1}^{M}P_{ik}^{\alpha_k-1}\right]\left[\prod_{i=1}^{N}\prod_{j=1}^{\rho_i}\sum_{k=1}^{M}P_{ik}q_{ijk}\right] \quad (6)$$

Taking the log, we get

$$\begin{aligned}log[L(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_N,\boldsymbol{\alpha};R)] &= NlogC(\boldsymbol{\alpha})+\sum_{k=1}^{M}(\alpha_k-1)\sum_{i=1}^{N}logP_{ik}\\ &+\sum_{i=1}^{N}\sum_{j=1}^{p_i}(log(\sum_{k=1}^{M}P_{ik}q_{ijk}))\end{aligned} \quad (7)$$

Multi-Genome Multi-Read (MGMR) algorithm

We wish to estimate $\boldsymbol{\alpha}$ and $\mathbf{p_1},\ldots,\mathbf{p_n}$ to maximize equation (7) above. For this purpose, we adopt an alternating iterative procedure of estimating $\boldsymbol{\alpha}$ given the current estimate of $\mathbf{p_1},\ldots,\mathbf{p_n}$ and vice-versa until the total change in $\boldsymbol{\alpha}$ becomes sufficiently small (or until a pre-set number of iterations have been executed).

Although for EM-based estimation methods convexity guarantees an optimal solution will be obtained, here (as shall be seen below) we have no such guarantee. Thus, we confine updates to be local by performing only one update for P and one for $\boldsymbol{\alpha}$. By one MGMR iteration, we refer to one EM-based P update followed by one $\boldsymbol{\alpha}$ update.

Estimating P given α

If we assume $\boldsymbol{\alpha}$ is given, we can write the EM steps to find $\mathbf{p_1},\ldots,\mathbf{p_N}$:

E step Letting ***Match*** signify a matching between reads and regions, and *Match(j)* be the region from which read j originates, we get:

$$\begin{aligned}log[L(\mathbf{P},\alpha;\mathbf{R},\mathbf{Match})] &= NlogC(\alpha)+\sum_{k=1}^{M}(\alpha_k-1)\sum_{i=1}^{N}logP_{ik}\\ &+\sum_{i=1}^{N}\sum_{j=1}^{\rho_i}(log(P_{iMatch(j)}q_{ij_{Match(j)}}))\end{aligned} \quad (8)$$

which leads to

$$Q(\mathbf{P},\alpha|\mathbf{P}^{(t)},\alpha^{(t)})=E_{Match|R,\mathbf{P}^{(t)},\alpha^{(t)}}[log(L)] \quad (9)$$

$$\begin{aligned}&= NlogC(\alpha^{(t)})+\sum_{k=1}^{M}(\alpha_k^{(t)}-1)\sum_{i=1}^{N}logP_{ik}\\ &+\sum_{i=1}^{N}\sum_{j=1}^{\rho_i}\sum_{k=1}^{M}(logP_{ik}+logq_{ijk})*\frac{P_{ik}^{(t)}q_{ijk}}{\sum_{k=1}^{M}P_{ik}^{(t)}q_{ijk}}\end{aligned} \quad (10)$$

$$\begin{aligned}&= NlogC(\alpha^{(t)})+\sum_{k=1}^{M}(\alpha_k^{(t)}-1)\sum_{i=1}^{N}logP_{ik}+\sum_{i=1}^{N}\sum_{j=1}^{\rho_i}\sum_{k=1}^{M}a_{ijk}logP_{ik}\\ &+\sum_{i=1}^{N}\sum_{j=1}^{\rho_i}\sum_{k=1}^{M}a_{ijk}logq_{ijk}\end{aligned} \quad (11)$$

where

$$a_{ijk}\equiv\frac{P_{ik}^{(t)}q_{ijk}}{\sum_{k=1}^{M}P_{ik}^{(t)}q_{ijk}} \quad (12)$$

M step Given that each q_{ijk} is fixed, the above reduces to maximizing

$$\begin{aligned}&NlogC(\alpha^{(t)})+\sum_{k=1}^{M}(\alpha_k^{(t)}-1)\sum_{i=1}^{N}logP_{ik}\\ &+\sum_{i=1}^{N}\sum_{j=1}^{\rho_i}\sum_{k=1}^{M}a_{ijk}*logP_{ik}\end{aligned} \quad (13)$$

Using Lagrange multipliers and differentiating, we see that this is maximized with

$$P_{ik}^{(t+1)}=\frac{\alpha_k^{(t)}-1+\sum_j a_{ijk}}{\sum_k(\alpha_k^{(t)}-1+\sum_j a_{ijk})} \quad (14)$$

Estimating α given P

Given a new estimate for $P^{(t)}$, we can use a fixed point iteration [12] to get a new estimate of $\boldsymbol{\alpha}$

$$\begin{aligned}F(\alpha) &= N[log\Gamma(\sum_k\alpha_k)-\sum_k log\Gamma(\alpha_k)]\\ &+\sum_{k=1}^{M}(\alpha_k-1)\sum_{i=1}^{N}logP_{ik}^{(t)}+Const(P^{(t)})\end{aligned} \quad (15)$$

By using the known bound $\Gamma(x)\geq\Gamma(\hat{x})exp((x-\hat{x})\Psi(x))$ $(\text{having}\Psi(x)=\frac{dlog\Gamma(x)}{dx})$, we can get a lower bound on F(α):

$$\begin{aligned}F(\alpha)\geq N[(\sum_k\alpha_k)\Psi(\sum_k\alpha_k^{(t)})-\sum_{k=1}^{M}log\Gamma(\alpha_k)\\ +\sum_{k=1}^{M}\alpha_k log\bar{P}_k^{(t)}+Const(P^{(t)})\end{aligned} \quad (16)$$

where $log\bar{P}_k=\frac{1}{N}*\sum_i logP_{ik}$.

We maximize this bound with a fixed point iteration similar to EM, noting that for fixed values of P convergence is guaranteed, and that for the Dirichlet distribution the maximum is the only stationary point [12]. This leads to the update

$$\alpha_k^{(t+)}=\Psi^{-1}[\Psi(\sum\alpha_k^{(t)})+log\bar{P}_k^{(t)}] \quad (17)$$

Heuristics/Implementation

As we have found $F(\boldsymbol{\alpha})$ presented in equation (15) is non-convex even in 2 dimensions (Figure 1), we confine updates to be local by allowing only one update for both the α and P estimation steps at each MGMR

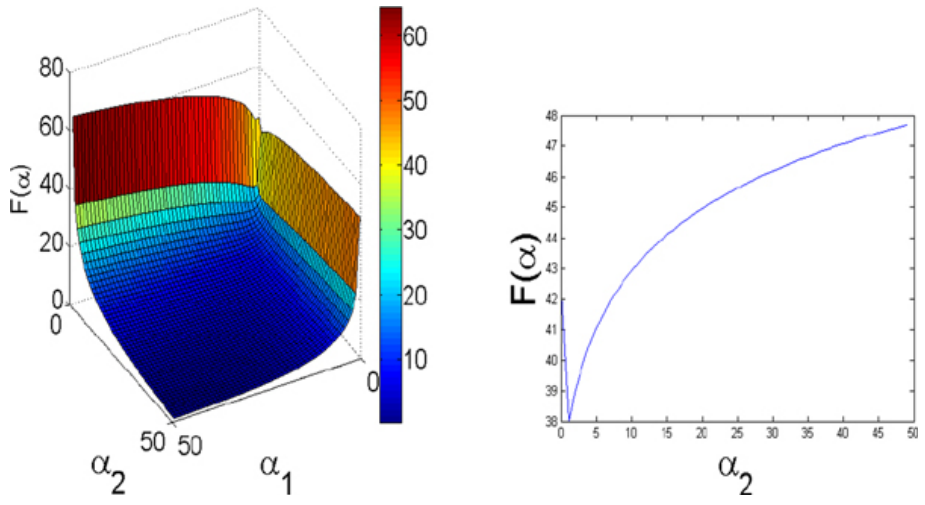

Figure 1 A mesh representation of F(α) [equation (15)] showing non-convex behavior. *P* is a 10 × 2 constant matrix and α is varied on [0:50,0:50]. The case shown is for N = 10, M = 2 (ten samples, two α parameters). Non-convex behavior is demonstrated by the values on the plane defined by α_1 = .06 on the range [0,50] on the right.

iteration. For genes with EM expression estimates equaling zero in all samples we substitute 10^{-20} for their values in MGMR to avoid taking the log of zero. For P updates (e.g., equation 14), we avoid potentially negative P values by adding one to each alpha (thus ignoring -1 in the numerator and denominator). We implemented the algorithm in MATLAB, where the inputs required are read-gene map files for each sample as in SEQEM [7], and an initial P estimate matrix. Alphas are initialized as an M-length vector of ones.

Results

Experimental setup

As in [7,9,10], we examined MGMR's accuracy by comparing its estimates of known expression levels with those of existing methods, namely SEQEM [7] and RSEM [9,10]. The initial "known" expression levels were estimates obtained from RNA-Seq samples; how these estimates were obtained is described below. In our case, we had to simulate a population sharing similar expression levels and the same set of gene regions. Our experiment differed in that we sought to use additional information to improve on the estimates of these existing methods. These methods were designed to estimate expression of single samples, and each had specific advantages which we disregarded in our comparison. For example, we ignored both SEQEM's ability to utilize SNP information and RSEM's ability to allow estimation on assembled transcripts by using only reference sequences.

Simulating data

To derive artificial reads, we first estimated expression on real biological samples using one method and then used the resulting distribution of expression values to generate simulated datasets for testing. Real samples were drawn from lymphoblastoid cells of the Yoruba in Ibadan (YRI) population [2,3]. As MGMR requires initial expression estimates, the estimate derived from the method it was being compared with in each case was input to MGMR. Thus, the four initial estimates used were from SEQEM, MGMR(SEQEM) (namely, MGMR initialized by SEQEM's result), RSEM, and MGMR (RSEM) (namely, MGMR initialized by RSEM estimates). We denote these four estimates SEQEM-A, SEQEM-B, MGMR-A and MGMR-B, respectively. We simulated reads by randomly selecting start sites falling within gene boundaries and extracting sequences from those positions. Read lengths were defined for each experiment, and simulations were repeated multiple times to account for randomness in the sampling process.

To derive the sequence set for the SEQEM comparison, we expanded upon the procedure used in [7]. There, SEQEM was shown to improve estimation of paralogous gene expression on a set of exon sequences from 51 Homo Sapiens chromosome 1 paralogs from the HomoloGene [13] database. We extracted a larger set of sequences containing all HomoloGene paralogs in Homo Sapiens having at least one exon longer than twice the read length used that do not overlap in genomic coordinates. We required this minimal length because sequences were sampled from exons, and we needed to ensure enough positions existed for full length reads to be sampled from these exons. 285 such genes remained (for reads of length 35 bp), and these were the genes on which expression was tested and

from which read sequences were derived. The SEQEM-A and SEQEM-B read sets were generated based on randomly selected exons from these genes and the expression levels from the SEQEM-A and SEQEM-B estimates taken on 20 YRI individuals. The read length of 35 bp corresponded to that of the YRI samples, and a coverage level of 20 was chosen, as this was the level at which SEQEM was shown to perform best in [7]. We performed a total of 30 repetitions of read simulations, where each repetition consisted of 20 samples (corresponding to the original 20 YRI samples used).

For the RSEM-A and RSEM-B read sets, the transcript set used was also obtained by filtering the HomoloGene database to avoid gene overlaps, but no length filtering was required: reads were now sampled directly from transcripts which all had effective lengths greater than the read length used. 524 transcripts corresponding to 265 genes survived this filtering. For these read sets, we produced 30 repetitions of 74 samples, where each consisted of 100 bp reads at a coverage level of 20. In all other respects the sampling process and read generation steps were identical to those performed for the SEQEM-A and SEQEM-B read sets.

Error measures

Accuracy was assessed by three error measures, the first two of which were applied in [7]: error rate, computed as $\frac{1}{n}\sum_i \frac{|P_i - Q_i|}{Q_i}$, χ^2 difference, computed as $\sum_i P_i \log \frac{P_i}{Q_i}$, and Kullback-Leibler divergence, computed as $\sum_i P_i \log \frac{P_i}{Q_i}$. Here P is the estimated distribution generated by the tested algorithm and Q is the true distribution. Error measures were averaged over all repetitions per sample, and then over all samples.

Simulated data with priors based on real estimates - estimating paralogous gene expression

To test performance on paralogous gene estimates, we set out to compare independent sample SEQEM estimates with MGMR's common population estimates. Before applying SEQEM, we looked to address one criticism of it from [11], where it was suggested that SEQEM's estimation could be improved by incorporation of transcript length correction. Upon examination, we found the effect of this correction was an increase in accuracy, and thus we maintained it for subsequent tests. This improvement is documented in the appendix.

With this correction in place, we estimated expression levels on the SEQEM-A and SEQEM-B read sets, applying SEQEM and MGMR to each. Outputs were recorded at 1-10, 20, 30, 40, 50 and 100 iterations for MGMR and at 100 iterations for SEQEM. The results are shown in Figure 2. We observed that both error and variance levels dropped sharply within just a few iterations for MGMR, and converged to significantly better estimates on average than SEQEM. These trends were consistent across all three error measures [Table 1]. Variance seemed to diminish more with MGMR over time, as might be expected for a method that shares information across samples. Notably, MGMR

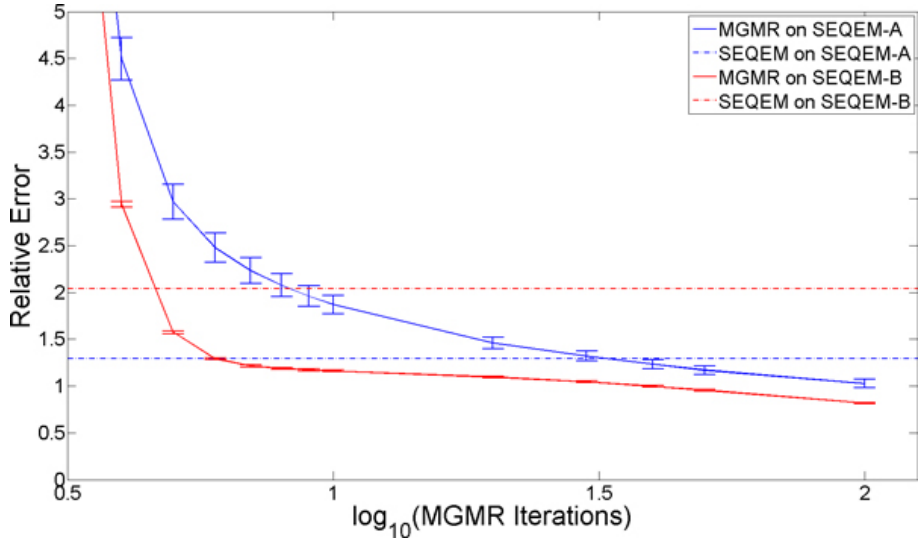

Figure 2 Relative error measured on SEQEM-A and SEQEM-B data sets. MGMR outputs on SEQEM-A and SEQEM-B initializations were compared with SEQEM up to 100 iterations. MGMR outputs were recorded at 1-10, 20, 30, 40, 50 and 100 iterations. The first few iterations have been trimmed to allow a compact presentation.

Table 1 MGMR vs. SEQEM error at 100 iterations on SEQEM-A and SEQEM-B data sets

	SEQEM-A sampling				SEQEM-B sampling			
	SEQEM		MGMR		SEQEM		MGMR	
	Error	SD	Error	SD	Error	SD	Error	SD
E	1.27	$1 * 10^{-2}$	1.03	0.14	1.50	0.70	0.82	$6 * 10^{-3}$
χ^2	0.66	$2 * 10^{-3}$	0.22	$4 * 10^{-3}$	0.69	0.05	0.27	$1 * 10^{-4}$
KL	0.29	$7 * 10^{-4}$	0.14	$1 * 10^{-4}$	0.18	$2 * 10^{-4}$	0.17	$1 * 10^{-4}$

These data sets were derived from SEQEM and MGMR(SEQEM) estimates, respectively, on 20 YRI samples. (E: relative error rate; χ^2: Chi-squared error; KL: Kullback-Liebler divergence; SD: standard deviation)

outperformed SEQEM on estimates for samples based on SEQEM-A, where sample estimates were obtained independently (and thus we expect the variation inherent in the real samples to be maintained).

Simulated data with priors based on real samples - estimating transcript level expression

We also sought to examine whether MGMR can improve results in the more challenging setting of estimating transcript level expression. Here, we expect estimates to be noisier due to low expression values in the real samples, and we must contend with multiread mappings due to paralogous genes as well as to isoforms of particular genes sharing subsequences as a result of alternative splicing. In anticipation of this challenge, we used a larger set consisting of 74 sample of single-end YRI samples as the real data source and simulated 100 bp reads instead of 35 bp. This was expected to be a difficult case for estimation, as all genes in the set are paralogs and many have multiple isoforms, as described in the section "Simulating data."

Once expression estimation was performed on the YRI samples and read sets RSEM-A and RSEM-B were generated, we again performed expression estimates with RSEM and MGMR on each set. In this case, unfortunately, we found the results did not exhibit a consistent trend as before and overall appeared inconclusive. These results are summarized in Table 2. It remains to be seen why the error results differ according to the level of estimation (gene vs. transcript) performed.

Table 2 MGMR vs. RSEM error at 100 iterations on RSEM-A and RSEM-B data sets

	RSEM-A Sampling				RSEM-B Sampling			
	RSEM		MGMR		RSEM		MGMR	
	Error	SD	Error	SD	Error	SD	Error	SD
E	0.1	$1 * 10^{-3}$	0.69	$1 * 10^{-3}$	1.0	$1 * 10^{-4}$	0.61	$1 * 10^{-3}$
χ^2	0.02	$6 * 10^{-4}$	1.25	0.01	0.02	$9 * 10^{-4}$	0.58	$3 * 10^{-4}$
KL	1.5	0.22	0.6	$1 * 10^{-3}$	0.8	0.11	0.38	$6 * 10^{-4}$

E: relative error rate; χ^2: Chi-squared error; KL: Kullback-Liebler divergence; SD: standard deviation

Table 3 Proportion of genes for which MGMR improves estimates on different data sets

	SEQEM-A	SEQEM-B	RSEM-A	RSEM-B
Proportion	104/285	78/285	126/524	173/524
%	36.5	27.3	24.0	33.0

Proportions of regions (genes for SEQEM and transcripts for RSEM, respectively) for which MGMR has lower relative error on average than each method compared to.

Conclusion

As shown by the 1000 Genomes and HapMap projects, one of the drives of modern genetics and bioinformatics research is to characterize variation in populations. Because of cost and time constraints, such projects have only recently become feasible. In addition to such studies assessing genomic variation and its relation to disease phenotypes based on DNA, it is anticipated that RNA-Seq population studies will also grow in popularity to more directly assign functional significance to variant loci by means of transcription measures. Thus, it becomes essential to accurately measure the expression levels from each individual to characterize such variation. Here, we have shown that for one common study design an unexpected benefit can arise. When individuals in these studies are drawn from the same population, the estimates made on each can be made more accurate because of the commonalities among population members.

A shortcoming of the MGMR approach is that since it assumes commonality among the samples, outlier samples will be attracted towards the common denominator, and thus appear more similar to the group profile than they really are. In particular, if the data are subject to differential expression analysis, MGMR may reduce the number of differentially expressed genes.

We have investigated the efficacy of MGMR in tackling two typical experimental settings - inferring expression levels of paralogs at the gene level, and of isoforms (also drawn from a difficult set of paralogs). Although substantial gains were obtained in the first, more inquiry is required to demonstrate a benefit in the latter. It is worth noting that in each case at least a quarter of the regions considered showed improvement, as shown in Table 3. With these results, we submit a proof of concept that population structure can aid in estimation of expression levels for RNA-Seq samples.

List of abbreviations

E: relative error rate; χ^2: Chi-squared error; KL: Kullback-Liebler divergence; SD: standard deviation; bp: base pair

Acknowledgements

E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. R.S. was supported in part by the European Community's

Seventh Framework Programme (grant HEALTH-F4-2009-223575 for the TRIREME project) and by the Israel Science Foundation (grant 802/08).
This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 6, 2012: Proceedings of the Second Annual RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-seq 2012).

Author details
[1]The Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel. [2]Molecular Microbiology and Biotechnology Department, Tel-Aviv University, Tel Aviv 69978, Israel. [3]International Computer Science Institute, Berkeley, CA, 94704, USA.

Authors' contributions
RR and EH developed the method. RS and EH designed the experiments. RR implemented the method and performed experiments. RR, EH, and RS analyzed results and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Published: 19 April 2012

References

1. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464**:773-777.
2. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768-772.
3. Frazer KA et al.: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851-861.
4. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11**:R25.
5. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics* 2010, **11**:94.
6. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology.** *Biol Direct* 2009, **4**:14.
7. Pasaniuc B, Zaitlen N, Halperin E: **Accurate estimation of expression levels of homologous genes in RNA-seq experiments.** *J Comput Biol* 2011, **18**:459-468.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**:511-515.
9. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty.** *Bioinformatics* 2010, **26**:493-500.
10. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12**:323.
11. Pachter L: **Models for transcript quantification from RNA-Seq.** *ArXiv e-prints* 2011.
12. Minka TP: **Estimating a Dirichlet distribution.** 2003 [http://research.microsoft.com/\~minka].
13. [http://www.ncbi.nlm.nih.gov/homologene/].

doi:10.1186/1471-2105-13-S6-S2
Cite this article as: Rozov *et al.*: MGMR: leveraging RNA-Seq population data to optimize expression estimation. *BMC Bioinformatics* 2012 **13** (Suppl 6):S2.