RECOUNT: EXPECTATION MAXIMIZATION BASED ERROR CORRECTION TOOL FOR NEXT GENERATION SEQUENCING DATA

EDWARD WIJAYA[1] MARTIN C. FRITH[1]
e-wijaya@aist.go.jp m.frith@aist.go.jp
YUTAKA SUZUKI [2] PAUL HORTON[1]
ysuzuki@k.u-tokyo.ac.jp horton-p@aist.go.jp

[1] *AIST, Computational Biology Research Center, 2-42 Aomi, Koutou-Ku, Tokyo 135-0064, Japan*
[2] *Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan*

Next generation sequencing technologies enable rapid, large-scale production of sequence data sets. Unfortunately these technologies also have a non-neglible sequencing error rate, which biases their outputs by introducing false reads and reducing the quantity of the real reads. Although methods developed for SAGE data can reduce these false counts to a considerable degree, until now they have not been implemented in a scalable way. Recently, a program named FREC has been developed to address this problem for next generation sequencing data.

In this paper, we introduce RECOUNT, our implementation of an Expectation Maximization algorithm for tag count correction and compare it to FREC. Using both the reference genome and simulated data, we find that RECOUNT performs as well or better than FREC, while using much less memory (e.g. 5GB vs. 75GB). Furthermore, we report the first analysis of tag count correction with real data in the context of gene expression analysis. Our results show that tag count correction not only increases the number of mappable tags, but can make a real difference in the biological interpretation of next generation sequencing data. RECOUNT is an open-source C++ program available at `http://seq.cbrc.jp/recount`.

Keywords: next generation sequencing, transcriptomics, tag count correction, sequence analysis

1. Introduction

In recent years, DNA sequencing technology has leapt forward with the advent of next-generation sequencing technologies such as Illumina GA (*aka* Solexa), Roche's 454, and SOLiD [9]. For example in the field of transcriptome analysis, by obtaining tens of millions of short reads from transcript populations of interest, new sequencing technology is enabling transcripts to be measured with unprecented accuracy and resolution [8, 12, 19]. Similarly in the area of metagenomics, the output of these technologies allows us to directly examine the molecular blueprints of microbial communities and determine their genetic variation [15].

Nevertheless, the high-throughput reads from these next-generation sequencing technologies contain substantial bias, because of the error rates that range from 0.3% at the beginning of reads to 3.8% - 25% at the end of reads [6, 18]. These errors introduce bias by reducing the quantities of the real reads and introducing false reads. We observed the effect of these errors on Solexa mouse data and found that more than 46% of the reads failed to map to the genome. These errors confound the measurement of real, but lowly expressed transcripts, and therefore can significantly reduce the quality of the conclusions which can be drawn from the data.

Many methods have been developed for SAGE read count correction. Some filter tags by forcing them to match known transcripts [2] or have a minimal quality and abundance [11]. Another attempted to join low-abundance tags to their neighboring (one-mismatch) tags [23]. Akmaev [1] extended this approach with a neighbor tag pair based procedure to discard spurious tags. Colinge and Feger [4] suggested a method to solve for the set of read counts whose expected observed counts (after sequence errors) equals the given observed counts. Finally Beißbarth et.al [3] introduced an Expectation-Maximization (EM) algorithm to find a set of estimated true read counts with maximal likelihood given the observed counts.

Unfortunately the software tools designed for SAGE cannot be directly applied to next generation sequencer data. This is due to differences in the details and amount of the generated data. For example: 1) Solexa reads have length greater than 30bp while SAGE tags are usually either 10 or 17 bp long, 2) the number of reads generated by Solexa typically is 100 fold greater than SAGE.

Recently Qu, et.al [18] proposed a clustering approach (FREC) to reduce the sequencing error in next generation sequencing data. The method uses an iterative procedure to cluster the reads and performs a sequencing error test for each cluster to assess the reads' membership to the cluster. The estimated counts of the representative reads is inferred from the total frequency of reads inside the cluster.

In this article we describe a tool – RECOUNT – designed especially to correct biases resulting from sequencing error in Solexa's reads. It adopts the EM algorithm of [3] to estimate the true counts/expression of the reads. Unlike FREC, RECOUNT exhaustively estimates the true counts for all the reads without pruning reads with low abundance. Although the running time is almost twice that of FREC, RECOUNT is much more memory efficient, using 14 times less memory than FREC. It also yields a higher percentage of mapped reads than FREC on some datasets and significantly outperforms FREC in making fewer large tag count errors when applied to simulated datasets.

We have applied RECOUNT to novel Solexa reads from mouse embryo, *Beta vulgaris* transcriptomes, 5'-end SAGE and bacterial metagenomics reads. In total they comprise more than 117 million reads. Evaluation in these datasets shows that RECOUNT increases the number of mapped tags by up to 13.85% and application on metagenomic data exhibits RECOUNT's ability to reduce the number of falsely mapped reads to the wrong genome. Furthermore we demonstrate that RECOUNT can prevent reporting false but apparently significant read count changes in tags

which map to annotated genes and pseudogenes in the mouse genome. We also identify the particular sequencing errors which cause the false observed counts in these cases.

2. Materials and Methods

2.1. *Data*

In our experiments we use four data sets: 1) transcriptome data from mouse embryo from 4 time points - day 7, 11, 15 and 17 (details of these data sets can be found in the Supplementary Material). 2) *Beta vulgaris* transcriptome data. It contains reads with length 27bp. It consists of more than 2 million reads [6]. 3) 5'-end SAGE data from *D. melanogaster*. It consists of more than 8 million reads of length 25bp [18]. 4) Metagenomic data. We analyzed metagenomic data from [15], which is known to come from the genome of *E. coli* strain *K12-MG1655*. It contains more than 6.5 million reads of length 36.

2.2. *Definitions*

For clarity, we start by formally defining a few terms. A *tag* is the DNA sequence of a sequencing read. In this study, within any dataset all tags are of the same fixed length. A *true tag* is the sequence of the actual DNA, while the *observed tag* is the output of the sequencer. The *neighborhood* of a tag t is the set of tags which have a non-neglible probability of being observed when t is the true tag or vice versa. A typical working definition of the neighborhood of tag t is all of the tags within Hamming distance 1 of t. A *library* is the multiset of tag sequences observed from one biological sample, *e.g.* a day 11 mouse embryo.

2.3. *Conversion from Solexa to Phred Error Probability*

The quality score of a base call is usually described in terms of *error probability*, namely the probability that a given base call is wrong [7]. We convert the Solexa quality scores to the more standard Phred score.

Let sQ be the Solexa quality score, pQ be the Phred quality score and ε be the error probability of a base in a given read. The Phred quality score is described as: $pQ = -10 \cdot log(\varepsilon)/log(10)$. We use the following formula for converting Solexa quality scores into Phred quality scores:

$$pQ = \frac{10 \cdot log(1 + 10^{\frac{sQ}{10}})}{log(10)}$$

For each unique tag sequence t and each position p in t, RECOUNT adopts the average of the error probability over all reads of t as the error probability for position p of t. When computing tag neighbors, we assume that each of the 3 possible substitutions at a given position are equally likely. Thus the error probabilities

are tag and position specific but not base specific. To speed the calculation, our implementation ignores the possibility of more than d errors in any tag, where d is typically set to 1 or 2.

2.4. *Statistical Model*

The error rates described in the previous section denote the probability that a true read i generates an observed read j as α_{ij}. Let N be the total number of unique reads in a library. The number of observed counts of a read is denoted as n_i and the true count is denoted as m_i, for $i = 0, \ldots, N$.

In forming a probability model, we assume the true read counts follow a Poisson distribution, namely given a true proportion p_j of a tag j, the true count is m_j with probability:

$$\frac{e^{-p_j\lambda}(p_j\lambda)^{m_j}}{m_j!}$$

for a fixed λ.

We adopt the Expectation Maximization algorithm [3, 5] to calculate the true counts given the observed counts and sequencing error rate estimates. The parameters we want to estimate are p_j and λ. The loglikelihood function is given by:

$$\lambda + \sum_{j=1,\cdots,N} \hat{m}_j log(p_j\lambda)$$

The details of the EM algorithm are as follows:

(1) E-step: Compute the likelihood and expected count of a tag j given by:

$$\hat{m}_j = \sum_{i=1,\ldots,N} \left(\frac{\alpha_{ij} p_j}{\sum_{k=1,\ldots,N} \alpha_{ik} p_k} \right)$$

(2) M-step: Maximize the likelihood of the complete data given the expected values and re-calculate new estimates for the parameters: $\hat{\lambda} = \sum_{k=1,\ldots,N} \hat{m}_k$ and $\hat{p}_j = \hat{m}_j/n$, where n is the total read counts in the library.

We iterate these steps until the parameters converge. We initialize the expected values $\hat{m}_j$ with the observed count of read j.

2.5. *Tag Correction Evaluation*

We use genome mapping to evaluate the effectiveness of tag map correction. The assumption is that tags which map perfectly to the reference genome are far more likely to be correct than other tags. Note that RECOUNT corrects tags solely on the basis of the library, without the use of a reference genome.

The genome mapping experiments carried in the next sections were done primarily using LAST. LAST is a general-purpose local alignment tool, broadly similar to BLAST, but much more efficient for genome-scale datasets (`http://last.cbrc.jp/`). Although it is not specialized for tag mapping, it fulfilled our needs and we understand it well. In Supplementary Material we describe in detail the usage of LAST in our experiments.

3. Results

3.1. *Effects of RECOUNT Error Correction on the Number of Genome Mappable Tags*

First we analyze the performance of RECOUNT by examining the effect of read count correction on the number of tags which can be mapped to the reference genome. In this experiment we mapped the four mouse embryonic transcriptomes and *B. vulgaris* libraries using LAST [a].

Figure 1 shows the number of *mapped* reads before and after applying RECOUNT or FREC [b] on the results. In this experiment RECOUNT was run with 1-Hamming distance neighborhoods [c]. For each library, the results from both mapping tools showed a substantial increase of mapped reads. On average RECOUNT increases the number of mapped reads by 13.85% whereas FREC does so by 11.55%. For the *B. vulgaris* and *D. melanogaster* datasets on average RECOUNT increases the number of mapped tags by 4.75% and FREC by 3.98%.

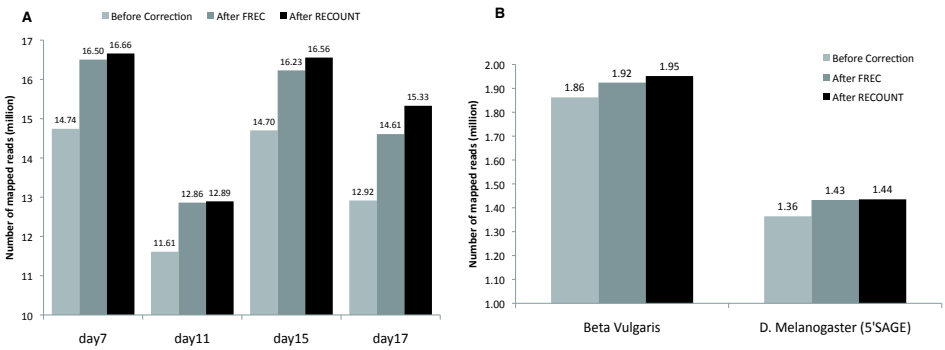

Fig. 1. Effect of RECOUNT error correction on the number of genome mappable reads. We also compare the performance of RECOUNT and FREC on A) mouse embryo and B) *B. vulgaris* and *D. melanogaster* (5'SAGE).

[a]Mapping on *D. Melanogaster* was done using ELAND as provided by Qu, et.al [18].
[b]We run FREC without using the adjusted quality value option. Hence the comparison between FREC and RECOUNT is based on the same error model.
[c]Experimental results on a small dataset (*B. vulgaris*) using 2-Hamming neighborhoods can be found in Supplementary Material.

Observe that although the performance of the two approaches is comparable, the advantage of RECOUNT over FREC is more evident when the total number of reads in the library increases (e. g. day 7, 15 and 17).

3.2. *Hamming Distance of Tags to the Genome*

In previous section, we used a heuristic mapping scheme that allows a few mismatches and indels: here we investigate genome matches in terms of Hamming distance.

We used LAST to divide the unique tags from the four mouse libraries into four categories (0, 1, 2, and ≥ 3) based on their Hamming distance to their best match in the genome. Figure 2A shows the changes of read counts before and after we apply RECOUNT. Notice the increase in read counts of perfectly matched tags (Hamming distance 0) after RECOUNT. Also, we see a significant decrease in the number of reads mapped with Hamming distance 1, 2 and ≥ 3 where the counts after RECOUNT become lower than before RECOUNT. This is because these counts have been carried over to the counts of reads with Hamming distance 0. Note that for all four cases the number of mappable tags given by FREC is also lower than RECOUNT. One of the advantage of RECOUNT is that total number of read counts before and after RECOUNT is applied are the same, however this is not the case for FREC.

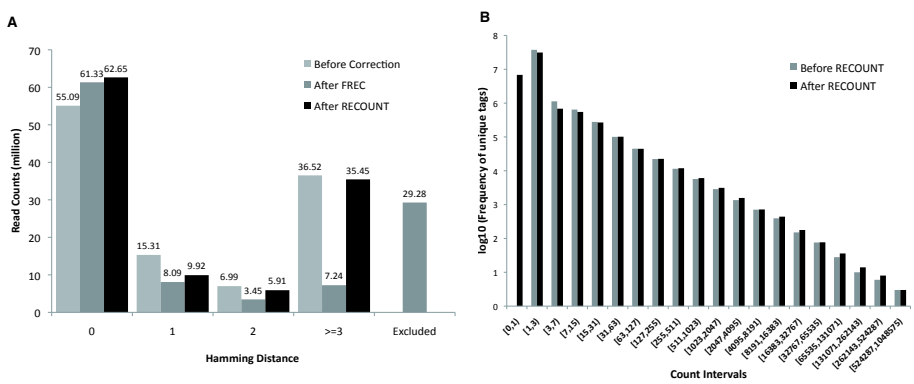

Fig. 2. A) RECOUNT reduced the number of mapped reads with mismatches and increased the count of perfectly matching reads. B) Change in count frequencies before and after RECOUNT. A histogram showing the number of unique tag sequences with read counts in each range, before and after applying RECOUNT is shown.

We also examined the frequency of the read counts before and after using RECOUNT on the four mouse libraries. We contend that if RECOUNT is effective, we should expect to see the number of unique tags with large counts increase and those with small counts decrease. This follows from the fact that tags with high read counts cannot be explained solely by random sequencing errors and thus tags

with high counts are likely to be real. Figure 2B shows the histogram of read counts before and after RECOUNT was applied. As expected, it shows that the frequency of large counts increases after the correction by absorbing counts from (apparently erroneous) small-count tags.

3.3. *Evaluation on a Simulated Data Set*

Evaluation based on the number of mapped tags may not give a full picture of the performance of RECOUNT. To further assess the effectiveness of RECOUNT we created a simulated data set in which we know in advance the number of true reads for each tag.

We constructed a pair of data sets: a *pre-simulated* and *post-simulated* library (refer to Supplementary Material for details). The pre-simulated library constitutes a library in which the tag count are *true*. We apply RECOUNT and FREC on the post-simulated library. The estimated counts from the post-simulated library are considered as *predicted counts*. The performance of the error correction tool then is measured based on the difference between the true counts and the predicted counts. Hence the lower the difference the better the performance.

Figure 3A below shows the frequency of tags based on the absolute difference. In general RECOUNT produces fewer tags than FREC with high absolute difference, and more tags with low absolute difference. For the tags with absolute difference [0,3) the total number of tags given by RECOUNT is 1.03 times more than FREC. For the tags with absolute difference [3,511) on average RECOUNT gives 4.47 times fewer tags than FREC. The most significant difference in performance happens in the range [7,15) where RECOUNT gives 6.69 times fewer tags than FREC.

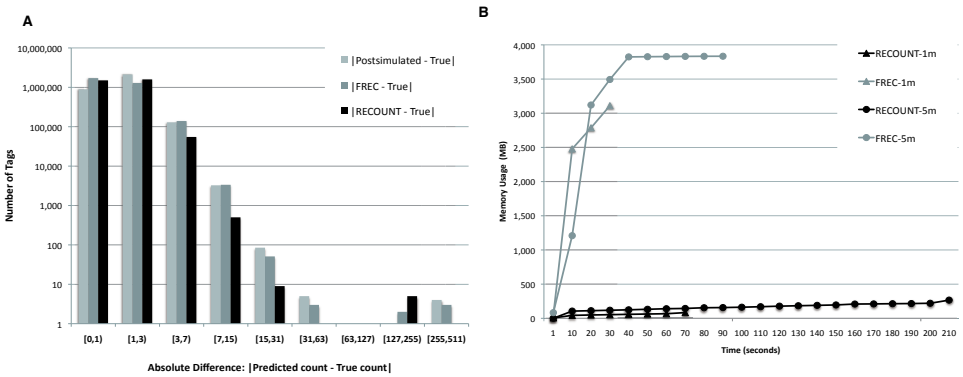

Fig. 3. A) Error correction performance on a simulated data set. B) Memory usage and running time of RECOUNT and FREC.

3.4. *Memory Usage and Running Time Comparison*

One of the crucial aspects in analysis of next generation sequencing data is the memory usage of the software. Since the data set sizes are growing much faster than computer memory sizes, there is a need for a tool that can effectively handle the massive output of the sequencer.

We compare the memory usage of RECOUNT and FREC using two subsets of mouse embryo data. The subsets contain 1MB and 5MB reads. Figure 3B shows that although on average RECOUNT is 1.35 times slower than FREC, the memory usage is 14.71 times less than FREC. When applied to the largest mouse embryo library (day 17), RECOUNT needs approximately 5GB of memory whereas FREC requires approximately 75GB.

3.5. *Analysis of Mapped Tags with Large Read Count Corrections*

Because of sequencing error, non-existing reads can be observed, and the read counts of true tags can be substantially altered. In the mouse embryonic data set we set out to investigate if we could detect such artifacts and determine if the correction done by RECOUNT affects the expression of known genes. For this purpose, for all of the mapped tags in the four mouse libraries, we found the corresponding mouse genes based on annotation in AceView [22].

We considered a tag to correspond to a gene if it mapped to within 500bp upstream or downstream from the transcription start site (TSS). For these experiments we allowed up to two mismatches when mapping to the genome. We compiled a list of tags from all the libraries where the read count change is greater than 50 fold after correction. Table 1 reveals that the read count of tags which correspond to *Hba*, *Dmkn*, and *Fabp1* have been substantially altered because of sequencing errors. In the observed data, the counts of these reads is lowered and the counts of their neighboring tags raised, both at substantial rates. As mentioned elsewhere in this manuscript, "neighboring tag" refers to the Hamming distance of the tags, not their genomic position, however note that in three of the four cases shown here, the neighboring tag shown maps to the same gene. In this case, the uncorrected data would not affect the estimated expression of the gene. The fourth example in Table 1 shows a case in which the neighboring tag maps to a different gene. The count of a read that corresponds to *EG408196* was reduced to zero, whereas its neighboring read that corresponds to *Stfa1* increased.

All of the genes mentioned in this section are potentially important: *Hba* expression was found to change in early stages of mouse embryo development [24], *Dmkn* is a gene primarily expressed in skin epithelial tissues but also expressed in other tissues [14], *Fabp1* is known to affect the growth and differentiation of mouse embryonic stem cells [21], and *Stfa1* was reported to be responsible for controlling susceptibility to autoimmune disease [10].

Table 1. Artifact of sequencing error on known genes, obtained from the pooled 4 mouse embryo data. The underlined nucleotides are the bases of the neighboring tags where the mismatch happened.

Tags	Obs. Count	Est. True Count	Genes
ACTTCTGATTCTGACAGACTCAGGAAGAAATCAT	2.20e3	0	Hba
ACTTCTGATTCTGACAGACTCAGGAAGAAACCAT	4.44e6	5.0e6	Hba
GGAAAGCAGGGAAGTCTGGGAACAGAGAGAGAAC	0.20e3	0	Dmkn
GGAAAGCAGGGAAGTCTGGGAACAGAGAGAGAAG	0.12e6	0.14e6	Dmkn
AGGCAGAGCTGTTGTGGTCAGCTGTAGAAAGGAA	0.10e3	0	Fabp1
AGGCAGAGCTGTTGTGGTCAGCTGTGGAAAGGAA	0.72e6	0.86e6	Fabp1
ATCATTTCTTCTCAGTGTCCAAGCCAGCAAGGAA	130	0	EG408196
ATCATTTCTTCTCAGTGTCCAAGCCAGCAAAGAA	56,266	68,756	Stfa1

3.6. *Changes of Expression for Known Genes*

RECOUNT clearly makes a difference at the indivual tag level. To investigate if it can also make a practical difference in analysis at the gene level, we used the tags to measure gene expression as in the previous section, by counting the overall number of read counts that mapped to within 500bp of the TSS of each gene. Using the four libraries of mouse embryonic transcriptomes, we identified genes with significant change in expression before and after read count correction by RECOUNT. Table 2 shows the list of highly affected genes. Observe that large gene read count reduction happens often with pseudogenes (i.e. genes with prefix "LOC"). This shows that RECOUNT is effective in correcting the expression of pseudogenes which are known to be unexpressed. We also identified several genes of interest with reduced expression after read count correction: *Mt_ATP*, a gene that is responsible for generating ATP synthase in mitochondria [13], *SUI1*, a gene that suppress intitiator codon mutations [17], and *Upf3a*, a gene that encodes a protein that is part of a post-splicing multiprotein complex involved in mRNA nuclear export and mRNA surveillance [20]. (Supplementary Material depicts the choromosomal view of gene expression change after RECOUNT is applied).

3.7. *Reduction of Falsely Mapped Metagenomic Reads to Wrong Strains*

One of the primary challenges with regard to metagenomics is how to deal with large tag libraries from diverse, often uncharacterized, genomes. Despite the enormous amount of sequence data that has been generated and analyzed in the past few years, publicly available software to help the analysis of metagenomic data is remarkably scarce [16].

We analyzed metagenomic data from [15], which is known to come from the *K12-MG1655* strain of *E. coli*. We further mapped the data to 6 closely related genomes, namely: *E. fergusonii*, *E. coli_O127:H6_E2348/69*, *E. coli_536*, *E. coli_55989*, *E.*

Table 2. Changes of expression of known genes.

Library	Obs. Counts	Est. True Counts	Genes	Fold Change	Library	Obs. Counts	Est. True Counts	Genes	Fold Change
day 7	194	0.01	LOC435256	192.84	day 15	117	0.00	LOC435256	117.00
	71	0.26	LOC627792	56.30		65	0.00	LOC625311	65.00
	54	0.00	LOC100043266	54.00		69	0.43	LOC667487	48.12
	53	0.00	EG665900	52.89		45	0.00	LOC667073	45.00
	49	0.03	LOC667073	47.48		43	0.04	LOC668239	41.51
	47	0.00	LOC667422	47.00		35	0.00	EG665964	35.00
	39	0.00	EG632013	38.96		30	0.00	LOC100039258	30.00
	36	0.00	LOC654358	36.00		28	0.00	LOC625328	28.00
	34	0.01	LOC100043780	33.76		24	0.00	LOC623798	24.00
	53	0.68	LOC624889	31.62		22	0.00	EG665900	22.00
	88	1.88	LOC667487	30.60		22	0.00	LOC100043225	22.00
	39	0.38	EG547263	28.36		28	0.31	LOC627792	21.41
day 11	92	0.00	LOC625311	92.00		20	0.00	LOC624889	20.00
	47	0.01	LOC666620	46.44	day 17	130	0.00	LOC435256	130.00
	57	0.31	EG665900	43.61		38	0.13	LOC668472	33.57
	40	0.01	rehiyora	39.60		32	0.05	EG629732	30.45
	38	0.00	SUI1.1	38.00		54	0.82	EG665964	29.74
	37	0.00	LOC100039422	37.00		30	0.04	LOC625328	28.90
	51	0.44	LOC625467	35.32		30	0.04	LOC667073	28.85
	34	0.00	EG547263	34.00		26	0.00	LOC100043266	26.00
	33	0.00	Mt_ATP-synt_D.0	33.00		39	0.51	LOC667487	25.79
	32	0.00	EG665964	32.00		50	0.94	LOC625311	25.77
	31	0.00	EG632013	31.00		25	0.09	LOC674926	22.87
	199	5.61	LOC433546	30.11		43	0.90	LOC100039258	22.61
	39	0.43	Hmg1l1	27.25		22	0.00	Upf3a	22.00
	47	0.99	LOC667487	23.63		20	0.00	LOC668773	20.00
	29	0.23	LOC100039258	23.52					

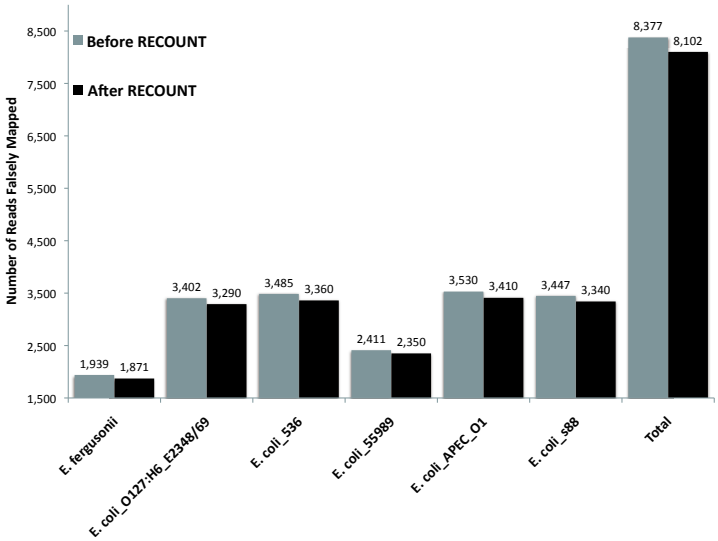

Fig. 4. RECOUNT reduced the number of reads misassigned to the wrong genome.

coli_APEC_O1, and *E. coli_s88*.

To examine if RECOUNT can reduce falsely mapped tags to the wrong strains, we looked at the number of reads that match *E. coli_K12-MG1655* with Hamming distance ≥ 1 but perfectly match another genome. Such mapping errors are relevant when judging if reads come from virulent microbes or closely-related but harmless microbes, for instance. Figure 4 showed that RECOUNT can reduce the number of falsely mapped reads in these wrong strains of *E. coli* by 3.29% in total.

4. Conclusion

In this article we have introduced a tool for correcting sequencing errors in next generation sequencing. We demonstrated the effectiveness of RECOUNT on several real datasets, showing that it can effectively decrease counts of false reads and increase the counts of true reads, as reflected by the significant increase of mapped tags. Compared with the recently published tool FREC [18], RECOUNT shows similar or better performance than FREC in terms of the number of genome mappable reads produced after read count correction. Application on simulated data set shows that RECOUNT significantly outperforms FREC in making fewer large tag count errors.

We also showed the effectiveness of RECOUNT in addressing real biological problems. For example the application of RECOUNT can have significant effects not only at the tag level, but also when tags are aggregated for gene level expression analysis. Examination of metagenomic data further shows RECOUNT does indeed reduce the number of reads falsely mapped to the wrong genomes; albeit

only slightly.

RECOUNT is scalable for Solexa reads. The running time for estimating the true counts from a library with 21 million Solexa reads is 4 hours on a 2.66GHz 64bit 8GB RAM Linux workstation. We believe that as next generation sequencers continue to improve they will generate more data. There is a great need for tools that can help biologists to interpret the transcriptomic data more accurately and effectively.

Acknowledgements

We thank Tim Beißbarth, Kentaro Tomii for their constructive comments and Wei Qu for her advice on running FREC. A part of the computation/mapping was done on the servers of the RNA informatics team in CBRC. This research was supported by a Japanese Ministry of Education, Culture, Sport, Science and Technology, Grant-in-Aid for Scientific Research (B) 19310128.

References

[1] Akmaev, V.R. and Wang, C. J, Correction of sequence-based artifacts in serial analysis of gene expression, *Bioinformatics*, 20:1254-1263, 2004.

[2] Bianchetti, L. *et al.*, SAGETTARIUS: a program to reduce the number of tag mapped to multiple transcripts and to plan SAGE sequencing tags, *Nucleic Acids Research*, 35(18):e122, 2007.

[3] Beißbarth, T., *et al.*, Statistical modeling of sequencing errors in SAGE libraries, *Bioinformatics*, 20:i31-i39, 2004.

[4] Colinge, J. and Feger, G. Detecting impact of sequencing errors on SAGE data, *Bioinformatics*, 17(9):840-842, 2001.

[5] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data using the EM algorithm. *Journal of Royal Statistical Society*, (39): 1-38, 1977.

[6] Dohm, J. C, *et al.* Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Research*, 36(16):e105, 2008.

[7] Ewing. B., and Green, P., Base-calling of automated sequencer traces using Phred. II. Error probabilities, *Genome Research*, (8):186-194, 1998.

[8] Fullwood, M.J., *et al.*, Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses, *Genome Research*, (19): 521-532, 2009.

[9] Metzker, M.L., Emerging technologies in DNA sequencing, *Genome Research*, (15):1767-1776, 2005.

[10] Mihelic, M. *et al.* Mouse stefins A1 and A2 (Stfa1 and Stfa2) differentiate between papain-like endo- and exopeptidases, *FEBS Lett.*, **580**(17), 4195-4199, 2006.

[11] Margulies, E. and Innis, J., eSAGE: managing and analyzing data generated with serial analysis of gene expression (SAGE), *Bioinformatics*, (16):650-651, 2008.

[12] Mortazavi, A., *et al.*, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, 5(7):621-628, 2008.

[13] Mootha, V.K, *et al.*, Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria, *Cell*, 115(5):629-40, 2003.

[14] Naso, M.F., *et al.*, Dermokine: An extensively differentially spliced gene expressed in epithelial cells, *Jour. of Inv. Dermatology*, (127):1622-1631, 2007.

[15] Nusbaum, C., *et al.* Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing, *Nature Methods*, 6(1):67-69, 2009.

[16] Pachter, L., Interpreting the unculturable majority, *Nat. Methods*, (4):479-480, 2009.
[17] Purothit, R., *et al*, Multiple translation initiation factor Sui1 related sequences in mammalian genomes, *Mamm. Genome*, 1(7):79-80, 2009.
[18] Qu, W., Hashimoto, S. and Morishita, S., Efficient frequency-based de novo short read clustering for error trimming in next-generation sequencing, *Genome Research*, (19):1309-1315, 2009.
[19] Shendure, J., *et al.*, The beginning of the end for microarrays?, *Nature Methods*, 5(7):585-587, 2008.
[20] Singh, G., *et al.*, Communication with the exon-junction complex and activation of nonsense-mediated decay by human Upf proteins occur in the cytoplasm, *Mol Cell.*, 27(5):780-92, 2007.
[21] Schroeder, F., *et al.*, Expression of liver fatty acid binding protein alters growth and differentiation of embryonic stem cells, *Moll. Cell. Biochem.*, 219(1-2):127-138, 2001.
[22] Thiery-Mieg, D., and Thiery-Mieg, J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation, *Genome Biology*, 7(Suppl 1):S12, 2006.
[23] Velculescu, V.E. *et al.*, Analysis of human transcriptomes, *Nature Genetics*, (270):484-487, 1999.
[24] Wilkinson, D.G. *et al.*, A molecular analysis of mouse development from 8 to 10 days post coitum detecs changes only in embryonic globin expression *Development*, 99(4):493-500, 1987.