BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 19 2009, pages 2613–2614
doi:10.1093/bioinformatics/btp438

Gene expression

Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite

Takehiro Hashimoto[1], Michiel J.L. de Hoon[1], Sean M. Grimmond[2], Carsten O. Daub[1], Yoshihide Hayashizaki[1] and Geoffrey J. Faulkner[2,3,*]

[1]Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan, [2]Expression Genomics Laboratory, Institute for Molecular Bioscience, University of Queensland, QLD, 4072, Australia and [3]Roslin Institute, University of Edinburgh, Roslin, Midlothian, EH25 9PS, Scotland, UK

Received on May 19, 2009; revised on June 20, 2009; accepted on July 13, 2009

Advance Access publication July 15, 2009

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: Multi-mapping sequence tags are a significant impediment to short-read sequencing platforms. These tags are routinely omitted from further analysis, leading to experimental bias and reduced coverage. Here, we present MuMRescueLite, a low-resource requirement version of the MuMRescue software that has been used by several next generation sequencing projects to probabilistically reincorporate multi-mapping tags into mapped short read data.

Availability and implementation: MuMRescueLite is written in Python; executables and documentation are available from http://genome.gsc.riken.jp/osc/english/software/.

Contact: geoff.faulkner@roslin.ed.ac.uk

1 INTRODUCTION

Next generation sequencing technologies have enabled high-throughput surveys of spatiotemporal expression across a broad range of biological contexts. The leading platforms at present generate millions of short (18–50 bp) reads per experiment. When applied to transcriptome and epigenome sequencing, using such techniques as shotgun sequencing of RNA (RNA-seq), chromatin immunoprecipitation sequencing (ChIP-seq) and Cap Analysis Gene Expression (CAGE) (Mortazavi *et al.*, 2008; Robertson *et al.*, 2007; Suzuki *et al.*, 2009), these short-read technologies present substantial bioinformatic challenges in mapping tags to a genome reference sequence.

One of the main problems with short-read sequencing is the substantial proportion of tags that map to multiple genomic loci. These multi-mapping tags (MuMs) are usually discarded from further analysis. This omission can introduce experimental bias, as MuMs provide information about transcribed genomic regions that cannot be detected with single map tags (SiMs) alone, such as active retrotransposons and gene families.

An alternative to the removal of MuMs is a strategy to assign them probabilistically to each genomic location to which they map. In a previous publication, we introduced a ‘guilt-by-association’ strategy to calculate the probability that a MuM originated from a particular locus (Faulkner *et al.*, 2008). MuMs were proportionately assigned to each of their mapping locations based on unique coincidences with SiMs and other MuMs. This MuMRescue algorithm was subsequently applied to large scale RNA-seq and CAGE data (Cloonan *et al.*, 2008; Faulkner *et al.*, 2009; Suzuki *et al.*, 2009), leading to substantially higher transcriptome coverage. During these implementations, we noticed that for some RNA-seq experiments, MuMRescue required >32 GB of RAM.

This observation was of critical importance considering the ever expanding throughput of next generation sequencing technologies. We subsequently aimed to increase the computational efficiency of the algorithm, benchmark it against other methods and make the software publicly available for the first time. Here, we present the result of this work, MuMRescueLite, an efficient ‘guilt-by-association’ rescue strategy for MuMs produced by large scale short-read sequencing experiments.

2 MuMRescueLite

The fundamental goal of MuMRescueLite is to calculate the probability that from a set of possible loci, a given locus is the true source of a MuM. This is achieved by counting the SiMs that occur in a nominal window around each locus occupied by a MuM and dividing this count by the total number of SiMs proximal to all loci associated with that MuM. In this way, MuMs are assigned proportionately to each of their loci and are therefore ‘rescued’. MuMs that do not coincide with at least one SiM are not rescued. Note that for RNA-seq data a window size of 200 bp is typically the point at which the proportion of MuMs rescued does not significantly increase with window size (and computational time).

MuMRescueLite is distinct from alternative methods such as splitting the signal associated with MuMs by the total number of locations to which they map (an equal weight approach), rescuing MuMs using both SiMs and other MuMs (as done by MuMRescue) or simply ignoring MuMs altogether. MuMRescueLite is also very different to the ERANGE approach, which assigns MuMs to known genes based on the SiM counts of those genes (Mortazavi *et al.*, 2008). To compare MuMRescueLite with these approaches, we benchmarked each in terms of computational requirements, percentage of MuMs rescued and RNA-seq/microarray correlations using publicly available data for mouse liver.

[*]To whom correspondence should be addressed.

Table 1. Comparison of MuM rescue methods

Rescue method	Maximum RAM (MB)	CPU time (s)	Percentage MuMs rescued	R (present on array)	R (>100 RPKM)
None	0	0	0.0	0.72 (11 247)	0.32 (686)
Equal weight	50	587	100.0	0.81 (11 247)	0.14 (760)
MuMRescueLite	476	3535	90.9	0.80 (11 247)	0.29 (756)
MuMRescue	10 652	27 527	93.5	0.79 (11 247)	0.3 (744)
ERANGE	2450	5742	91.6	0.81 (11 247)	0.16 (768)

A window size of 200 bp was used for both MuMRescueLite and MuMRescue. A total of 13 289 953 SiMs and 3 765 156 MuMs (mapping to 10 or fewer locations) were generated by the RNA-seq experiment (Mortazavi *et al*., 2008). Pearson correlation (R) values were calculated for RefSeq transcripts called as present based on the Affymetrix data (GEO Ref: GSE6850) or generating >100 reads per kilobase, per million mapped reads (RPKM) in the RNA-seq experiment The number of RefSeq transcripts used to calculate R is indicated in brackets. Benchmarking was performed on a 2600 MHz, dual core AMD Opteron™ processor with 64 GB of RAM.

As shown in Table 1, MuMRescueLite required more RAM and CPU time than equal weighting or not rescuing the MuM RNA-seq tags and less RAM and CPU time than the MuMRescue or ERANGE algorithms. MuMRescueLite, MuMRescue and ERANGE rescued a similar percentage of MuMs (∼91%) but obviously rescued fewer than equal weighting (100%).

Cross platform correlations between the RNA-seq and Affymetrix data were consistently highest for MuMRescueLite and MuMRescue. For this analysis, we calculated Pearson correlations (R) for RefSeq transcripts either called as present by the Affymetrix experiment or RefSeq transcripts reliably called as present by the RNA-seq experiment (Table 1). In the regards to the former, each method other than no MuM rescue generated a correlation of ∼0.8, with no MuM rescue leading to a correlation of 0.72. Conversely, for RefSeq transcripts called as present by the RNA-seq experiment, MuMRescue, MuMRescueLite and no MuM rescue generated correlations of ∼0.3, compared with equal weighting and ERANGE which lead to correlations of ∼0.15. These results suggest that MuMRescue and MuMRescueLite provide the highest rates of true positive assignments and the lowest rates of false positive assignments of MuMs to RefSeq transcripts.

Overall, the benchmarking demonstrated that MuMRescueLite provided the best combination of computational efficiency, success in rescuing MuMs and cross platform agreement. The very low RAM requirements of MuMRescueLite would permit much larger scale sequencing datasets to still be processed on a standard PC.

However, the merits of the other approaches should not be ignored. The ERANGE package provides a useful variation of the fundamental 'guilt-by-association' strategy pioneered by MuMRescue by using other evidence such as ESTs, cDNAs or known genes to resolve MuMs. However, the reliance of ERANGE upon predefined gene models is highly subject to the definition of what constitutes a 'gene'. This notion is increasingly complicated by pervasive intergenic transcription in mammals (Birney *et al*., 2007) and transcriptional complexity within mammalian genes.

By using a window around each MuM, MuMRescueLite avoids the need to define a gene set. Furthermore, it can be customized to use a window specific to a given genomic structure, such as proximal promoter regions (Carninci *et al*., 2006). Yet another positive aspect of the use of a window is that as sequencing depth increases, the probability of a MuM occurring near another tag also increases. This permits more MuMs to be rescued, as we observed for 'deep' CAGE sequencing (Suzuki *et al*., 2009), where far more CAGE tags were rescued as a proportion than for earlier CAGE experiments. Finally, the simplicity of use for MuMRescueLite, which requires a single input file, a window size and the output file name, compares well with ERANGE, which needs several input files and packages as part of a larger pipeline.

In a final note to users, MuMRescueLite is designed to analyze individual short read libraries corresponding to a specific tissue, cell or activation state, rather than combining multiple libraries of varying biology. If the latter is done, a SiM may rescue a MuM when in fact these tags originate from different biological states and therefore may not coincide *in vivo*.

3 CONCLUSIONS

MuMRescueLite is an evidence-based treatment of multi-mapping short sequence reads. It is as accurate as its precursor, MuMRescue, but is far more computationally efficient due to code optimization and the use of SiMs alone to rescue MuMs. We have shown that MuMRescueLite is superior to disregarding MuMs or equal weighting their signal. At least for the RNA-seq data presented here, it also compares favorably with ERANGE in several areas. Finally, MuMRescueLite greatly expands the coverage and mapping rate of massively parallel short read approaches to transcriptome and epigenome sequencing, with millions of informative tags 'rescued' and an increase in cross-platform correlation.

Funding: RIKEN Omics Science Center and the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan; Functional RNA Research Program (supported by the RIKEN Frontier Research System, to Y.H.); National Health and Medical Research Council (grant 455857 to S.M.G and grant 456140); Australian Stem Cell Centre (to G.J.F).

Conflict of Interest: none declared.

REFERENCES

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Carninci,P. *et al*. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

Cloonan,N., *et al*. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. methods*, **5**, 613–619.

Faulkner,G.J. *et al*. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, **91**, 281–288.

Faulkner,G.J. *et al*. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Suzuki,H., *et al*. (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.*, **41**, 553–562.