Effects of Incomplete Reference Transcriptomes on RNA-Seq Mapping for Higher Eukaryotes

Alexis Black Pyrkosz[1], Hans Cheng[1], C. Titus Brown[2,*]
1 Avian Disease Oncology Laboratory, USDA, East Lansing, MI, USA
2 Microbiology Department, Michigan State University, East Lansing, MI, USA
∗ E-mail: Corresponding ctb@msu.edu

Abstract

Introduction

Whole transcriptome sequencing using next-generation sequencing (NGS) technologies, or RNA sequencing (RNA-seq) has been growing in popularity for studying non model organisms in the fields of agriculture, evolution, and medicine (CITE). With NGS technology evolving to the point where the cost is within the budgetary range of many funding agencies and the resultant data can be mined for a range of applications (CITE), an increasing number of labs are choosing NGS and RNA-seq for their research. However, many of the NGS computational tools were developed for model organisms (E. coli, yeast, human...) (CITE) for which a complete genome and potentially a complete transcriptome is available. In this study, we use simulation techniques to assess whether current RNA-seq computational tools are applicable for organisms that do not have a complete transcriptome.

RNA-seq as a platform for exploring gene expression

Many genome sequencing projects are focused on identifying genome sequences that are transcribed into functional mRNAs that affect downstream biology (CITE). RNA-seq or cloning and sequencing of the cDNA is used for generating the transcriptome, a full set of transcripts that includes splicing isoforms. Unlike genomic sequencing where the entire genome is sequenced at a given level, RNA-seq measures the expression level of individual transcripts, which can differ by up to three orders of magnitude within the same cell (CITE). Differential gene expression has been used in a variety of different organisms, tissues, environmental conditions, disease states, etc to determine which genes/transcripts are up or down regulated between two samples. While microarrays have traditionally served this same purpose, RNA-seq is superseding due to greater and more accurate information.

An RNA-seq experiment begins by converting mRNA into a library of randomly fragmented cDNA. Then one of the NGS technologies are applied: Illumina, Roche 454, or SOLiD. Sequencing generates millions or billions of short reads either from one end of the cDNA fragments (single-end reads) or from both ends (paired-end reads). The raw data consist of several gigabytes of short sequences with associated quality scores. Illumina is the most common NGS technology owing to the higher number of reads produced at cost, despite being known for having a substitution error rate on the order of 1% and single-end reads with a current maximum length of 100 bp. Roche 454 offers reads between 400 and 500 bps, but with significantly lower sequencing depth for the same cost. SOLiD ... Because Illumina is the most widely used of these platforms, our simulations use Illumina-style reads for testing the RNA-seq computational tools.

Computational methodology for RNA-seq

While it has been shown that high sequencing depth is required for detecting low abundance transcripts and isoforms, making it desirable to obtain a high sequencing depth (CITE), the sheer amount of data (millions to billions of reads) generated during sequencing complicates downstream computational analyses. The crucial step in converting the reads to useful sequencing information is mapping the reads to a reference transcriptome. This reference is either a standard available from online databases such as Ensembl (CITE) or is built through de novo assembly of the reads. Several NGS read mapping and alignment programs are available and most of them use either a fast indexing algorithm to quickly identify potential matches to the reference or an index based on counting short k-mers that are in both the read and the reference (CITE). Most mappers are designed to allow mismatches between the read and the reference, owing to the substitution errors and indels that can result from the various sequencing platforms. For large data sets, a tradeoff between speed/memory and accuracy must be made and most mappers contain parameters that users can tweak as appropriate for their data.

One parameter of importance for the current work addresses the multi mapping conundrum. Multimap reads are reads that map equally well to several locations in the transcriptome. The default for most mappers is to randomly select a position from those that match the read and report that position only. Some mappers will always report the first position found as the matched position while others can be programmed to find a 'best' match according to a specified criterion. This default results in rapid mapping, but also many erroneous matches that can lead to artificially higher/lower transcript expression levels further down the computational pipeline. The other two common settings for this parameter are unique and multi mapping. For the unique parameter, if a read matches equally well to two or more positions, then it is simply discarded. While this prevents false positives from being reported, in the case of two highly similar isoforms, nearly all of the reads will be discarded, resulting in an artificially low expression level for both transcripts. For the multi mapping parameter, all possible matches are reported. While this guarantees that the correct location is reported, it also lists many incorrect positions that must be filtered out by another metric. There have been several attempts to statistically allocate or distribute the multi map reads with varying levels of success (CITE). It has been predicted that use of longer reads or paired-end reads is the ultimate way of solving this problem. In this study, we show that longer reads are indeed a solution to multi map problem, but technology must advance to produce significantly longer reads than are currently available before the high false positive rates decrease to negligible levels.

Underlying assumptions in RNA-seq computation that are not necessarily valid for all organisms

The primary assumption made in current mappers is that the reference transcriptome is complete. For model organisms used during development of these tools (human, mouse, yeast), the transcriptome has been built from the genome and the genome frequently has Sanger sequencing, BACs, hybrid and radiation maps (and what else....) forming the skeleton of the reference with NGS sequencing information incorporated. For these organisms, the reference has been extensively refined and is available from online databases. For many non model organisms, there is no reference available and the cost of obtaining physical/genetic maps is prohibitively high (CITE). Researchers working on these organisms create their own reference using de novo assembly.

Constructing a de novo assembly requires overlapping short, potentially low quality reads into contigs and then linking these into supercontigs (CITE). If Sanger sequencing or other prior information about the chromosome content is available, scaffolding can be used to further piece the supercontigs together (CITE). Transcriptome assembly is more challenging than genome assembly owing to the varying sequence depth of the individual transcripts, strand specificity, and isoforms (CITE). These challenges are less of an issue with bacteria, archaea, and lower eucarya because highly repetitive sequences and alternative splicing is limited (less than X%). However, higher eucaryal genomes can have significant repetition (up to X% in species) and up to X% isoforms, which the various assemblers are presently still struggling to resolve. For this reason, many reference transcriptomes based solely on NGS data are incomplete and contain misassemblies. Also, assembling a de novo transcriptome is a nontrivial task with potentially more than 128 GB of RAM and more than a week of computational time on supercomputing clusters (or the cloud) required. While efforts are underway to improve assembly efficiency, the problem remains that the biology for higher eucarya is more complex.

Complexity of higher eukaryotic genomes

There are two primary areas of complexity that affect de novo transcriptome assembly in higher eucarya: sequence repeats and alternative splice variants. Sequence repeats... Alternative splicing ...

Goals of this study

In this study, we use simulation to examine the effects of incomplete reference transcriptomes on RNA-seq mapping. We first characterize the false positive rates that are occur when mapping reads to simulated and real transcriptomes, thereby showing how isoforms cause significant error in mapping, which increases as the reference is less complete. We perform mapping with three common mapping programs and show that isoforms are a problem for all of them, indicating that the isoform problem is a general issue. We then suggest that clustering transcripts likely to be isoforms is a means of temporarily handling the isoform problem for mapping. We finish by showing that increasing the length of the reads or using paired end reads will greatly improve the false positive rate due to isoforms, but not completely resolve it.

Results

Isoforms from real transcriptomes are the primary source of false positives

False positive rates from mapping increase as completeness of reference decreases

Choice of mapping programs has little effect on mapping accuracy

Transcript families are a means of grouping isoforms for further data analyses

Reads must be considerably longer to resolve the isoform problem

Current paired end read lengths are insufficient to resolve isoforms

Discussion

Random data is not a good substitute for real sequencing data

More sequencing data will not solve the isoform problem

Isoforms will most likely be resolved by additional experiment

New technology is needed to decrease false positive rates

Conclusion

Materials and Methods

Simulation of random transcriptomes

Simulation of isoforms

Simulation of reads

Mapping

Read mapping accuracy

Acknowledgments

References

Figure Legends

Tables

Table 1. Effect of 1% Substitution Error on Mapping of Simulated Reads to Random and Real Transcriptomes

Organism	Num Trans	Error	TP (d)	FP (d)	TP (u)	FP (u)	TP (m)	FP (m)
Random	5000	1%	92%	0%	92%	0%	92%	0%
Mouse	5000	1%	87%	5%	81%	0%	92%	12%
0% Isoforms	5000	1%	92%	0%	92%	0%	92%	0%
10% Isoforms	5000	1%	86%	6%	81%	0%	92%	14%
20% Isoforms	5000	1%	80%	12%	70%	0%	92%	33%
30% Isoforms	5000	1%	74%	18%	62%	0%	92%	60%
40% Isoforms	5000	1%	67%	25%	53%	0%	92%	105%
50% Isoforms	5000	1%	60%	32%	45%	0%	92%	217%
60% Isoforms	5000	1%	55%	37%	39%	0%	92%	303%
70% Isoforms	5000	1%	48%	45%	33%	0%	92%	732%
80% Isoforms	5000	1%	42%	50%	27%	0%	92%	970%
90% Isoforms	5000	1%	37%	56%	23%	0%	92%	4722%

The random transcriptome shows the case where transcripts are all unique (no isoforms) and therefore in the absence of error, all reads will be mapped to their original transcripts. Addition of a 1% error will cause 8% of the reads to be discarded because they cannot be mapped to any transcript using normal parameters. However, in the mouse transcriptome that contains isoforms, 2% of the reads will be mapped incorrectly even in the absence of sequencing error. The 1000 transcripts were salected randomly. (d), (u), and (m) are default, unique, and multimap, respectively, indicating which parameters were used for the mapping. TP and FP are true positive and false positive as measured by a read being mapped to its original transcript or being mapped to a different transcript.

Table 2. Comparison of False Positive Rates Between Model and Nonmodel Organisms with 20X coverage and 1% substitution error

Organism	Num Trans	TP (d)	FP (d)	TP (u)	FP (u)	TP (m)	FP (m)
Homo sapiens (Human)	180223	38%	54%	21%	0%	%	%
Mus musculus (Mouse)	88078	47%	45%	26%	0%	92%	233%
Takifugu rubripes (Pufferfish)	47994	40%	52%	19%	0%	92%	267%
Rattus norvegicus (Rat)	34721	62%	30%	41%	0%	92%	101%
Caenorhabditis elegans (Worm)	31264	61%	31%	43%	0%	94%	133%
Ornithorhynchus anatinus (Platypus)	27380	62%	30%	37%	0%	92%	80%
Canis familiaris (Dog)	27251	75%	17%	61%	0%	92%	55%
Drosophila melanogaster (Fruitfly)	24612	46%	46%	28%	0%	95%	352%
Xenopus tropicalus (Frog)	22878	76%	16%	61%	0%	93%	71%
Sarcophilus harrisii (Tasmanian devil)	22582	76%	16%	61%	0%	92%	34%
Gallus gallus (Chicken)	22215	72%	20%	56%	0%	92%	68%
Ailuropodo melanoleuca (Panda)	21891	84%	8%	76%	0%	92%	18%
Taeniopygia guttata (Zebra Finch)	18560	86%	6%	79%	0%	93%	70%
Tursiops truncatus (Dolphin)	17523	83%	1%	83%	0%	84%	10%
Choloepus hoffmanni (Sloth)	14039	67%	1%	66%	0%	68%	3%
Petromyzon marinus (Lamprey)	11338	84%	8%	77%	0%	94%	23%
Saccharomyces cerevisae (Yeast)	6757	87%	6%	84%	0%	92%	86%

Table 3. Comparison of Accuracy

Organism	Num Trans Ref	Error	TPs (d)	FPs (d)	TPs (u)	FPs (u)	TP (m)	FP (m)
Mouse	88366	1%	47%	45%	27%	0%	92%	235%
Mouse	90%	1%	44%	45%	26%	2%	83%	212%
Mouse	80%	1%	43%	44%	25%	4%	74%	187%
Mouse	70%	1%	38%	44%	24%	7%	65%	165%
Mouse	60%	1%	34%	43%	22%	9%	55%	140%
Mouse	50%	1%	30%	41%	21%	11%	46%	118%
Mouse	40%	1%	26%	38%	19%	14%	37%	94%
Mouse	30%	1%	21%	33%	16%	16%	28%	69%
Mouse	20%	1%	15%	28%	13%	16%	19%	48%
Mouse	10%	1%	8%	16%	8%	13%	9%	23%
Chicken	22290	1%	72%	20%	56%	0%	92%	69%
Chicken	90%	1%	66%	20%	53%	2%	83%	62%
Chicken	80%	1%	60%	19%	49%	4%	74%	55%
Chicken	70%	1%	54%	18%	45%	6%	65%	48%
Chicken	60%	1%	48%	18%	41%	8%	56%	40%
Chicken	50%	1%	41%	16%	36%	9%	46%	34%
Chicken	40%	1%	33%	14%	29%	9%	36%	27%
Chicken	30%	1%	25%	12%	23%	8%	27%	21%
Chicken	20%	1%	18%	9%	17%	7%	19%	14%
Chicken	10%	1%	9%	5%	9%	4%	9%	7%

List of test cases and mapping accuracy. Num Transcripts Ref indicate the number of transcripts in the reference (reads were simulated for the complete reference) respectively. Error indicates the percent random substitution error in the reads. TP is the true positive or percentage of reads that were mapped to their original transcript. FPs indicates the false positives or reads that are reported to align but are mapped to the wrong transcript. (d), (u), and (m) are the default, unique, and multi map parameters used. Each simulation at the 50% transcriptome expression level was run in triplicate with the results averaged.

Table 4. Comparison of Three Common Mapping Programs on the Same Chicken Data Set

Organism	Num Trans	Error	Bowtie TP (d)	FP (d)	BWA TP (d)	FP (d)	SOAP2 TP (d)	FP (d)
Chicken	100%	1%	72%	20%	72%	20%	78%	22%
Chicken	90%	1%	66%	20%	66%	20%	72%	22%
Chicken	80%	1%	60%	20%	60%	19%	79%	19%
Chicken	70%	1%	54%	19%	54%	19%	58%	21%
Chicken	60%	1%	47%	18%	48%	18%	51%	20%
Chicken	50%	1%	41%	16%	41%	16%	44%	18%

Comparison of Bowtie, BWA, and SOAP2 mapping programs on the same simulated reads for the chicken data set with decreasing completeness of the reference transcriptome.

Table 5. Transcript Families

Table 6. Comparison of Effects of Longer Reads on Mapping to Complete Mouse Transcriptome with 20X coverage and 0% error

Read Length (bp)	TPs (d)	FPs (d)	TPs (u)	FPs (u)	TP (m)	FP (m)
100	51%	49%	28%	0%	100%	265%
200	55%	45%	32%	0%	100%	223%
300	58%	42%	36%	0%	100%	193%
400	60%	40%	39%	0%	100%	174%
500	63%	37%	42%	0%	100%	158%
600	65%	35%	45%	0%	100%	146%
700	66%	34%	47%	0%	100%	136%
800	68%	32%	49%	0%	100%	128%
900	69%	31%	51%	0%	100%	121%
1000	70%	30%	53%	0%	100%	114%