Survey: RNA-Seq Analysis Pipeline for Differential Gene/Transcript Expression Analysis

bodhisattvax@gmail.com

93 Responses as of 4th December 2012

Question 1: What do you prefer to align your reads to?

	Response Percent	Response Count
Genome	47.3%	44
Transcriptome	12.9%	12
Both (e.g. when you supply Tophat with both)	39.8%	37
If possible, please specify the reason for your preference Show Responses		25
answered question		93
skipped question		0

Reasons for choosing what to align against:

*** Genome***

"Often the only option i have available"

"The amount of intergenic, or intronic mapping is important in my interpretations."

"using the whole genome can allow the identification of novel features and not just those already known"

"Supply the Genome and use the transcriptome GTF"

"The genome is a unique reference while transcripts are redundant and overlap with each other"

"ability to discover new transcripts"

"Genome is more available"

"Identify alternative splicing, gene fusions and novel RNA transcripts (lincRNAs)"

"Unbiased alignment to take into account unannotated transcripts"

"not restricted to (only) one transcript assembly"

"Don't want to be restricted to existing annotation. Of course this could vary from project to project"

"More accurate/well annotated"

"I align to the genome to have the best chance of finding novel spliceforms."

Transcriptome

"No reference genome available"

Both(e.g. when you supply Tophat with both)

" interested in transcript variants"

"I use the tophat transcriptome then genome alignment method. I use this because, first since the transcriptome is a much smaller search space and will contain a large portion of reads, this will decrease the search time. Second, because if a read aligns to that region it makes more sense then if it aligns somewhere else, so by using this method it, in my mind, correctly solves the problem of when a read maps to multiple locations."

"I think it's good to use both if you have them, and in each case, it's nice to be able to align to one or the other."

"The main problem is pseudogenes, to really get better mapping you need to run in genome and transcriptome."

"I generally only use a transcriptome fasta file when estimating insert sizes. Otherwise I use the whole genome and a junctions file of some sort (GTF) from encode. I find this offers the most flexibility, especially if I also want to discover possible novel junctions."

"Aligning to genome helps with specificity and can align to locations that are outside your annotated transcripts. Aligning to transcriptome allows reads to span junctions (without having to rely on finding junctions de-novo)."

"Depends which reference you have available and what makes more sense for computational time or biological question."

"I'm doing a Phd using RNA seq but I've worked on genome analysis too."

"It depends very much on what I am doing - for some analysis it is more convenient to align against the transcriptome, for others it is absolutely necessary to use the genome. It is not a question that can be answered generally"

"If by "both" you mean giving Tophat a FastA genome sequence and then a GTF describing the transcriptome, then, sure, would won't prefer to have more information for Tophat to use? If I have an uncharacterized genome then I'll just go a de-novo assembly route."

"I find the tophat solution to be a pretty good tradeoff. Aligning only to the transcriptome has problems."

Question 2: What is your preferred aligner?

	Response Percent	Response Count
TopHat	67.9%	55
Bowtie(only)	17.3%	14
STAR	6.2%	5
BWA	4.9%	4
GSNAP	3.7%	3
	Other (please specify) Show Responses	16
	answered question	81
	skipped question	12

Other aligners (number of times mentioned) : segemehl(2), MapSplice(1), GEM(2), RUM(1), MIRA(1), NEWBLER(1), Lifescope(1), NovoAlign(2), Shrimp(1), Bioscope(1), X-mate(1), indexDP(1)

Question 3: Reasons for choice of aligner

Tophat

"ease of use"

"Easy. STAR has consistently failed to run. I'd like to try it for it's supposed speed increase and memory usage decrease."

"I chose tophat because it is the only truly gapped aligner meant for transcriptome analysis."

" it works "

"it is very popular and wildly used from tophat to cufflink"

"Tophat have been improved over time, with bowtie2 I will expect than mapping will be better too."

"Fast and accurate splice mapper"

"The cufflinks suite is nice. I.e. i have experience with it, and have not been forced to change yet ;)"

"To identify novel splice variants"

" ease of use ; documentation"

" TopHat -> popular, Newbler -> 454, friendly. MIRA -> popular and excellent mailing list"

"fast and accurate"

"Largely historic - we adopted it very early and a lot of projects are tied into using it so it is not convenient to switch to something else, moreover there is no evidence the alternatives are sufficiently better to justify such a move"

"easy to use, works well, no hassles"

"Tophat could align spliced reads"

"Tophat is still #1 rated. Bowtie, BWA won't handle introns. Star might be good but I haven't looked at it."

"Reliable, easy to get support"

"Gapped alignment allowing exon-junction determination of alternative transcripts."

"I've tried tophat, bwa, and gsnap (as well as novoalign, which wasn't listed), and found tophat (tophat2 with bowtie2) to have a good tradeoff between speed and precision. GSNAP is quite good as well for RNAseq data, I just find tophat a bit more convenient."

Bowtie(only)

"Ease of use."

"Trinity was used for transcriptome assembly, Bowtie is a built-in function in Trinity. It's also fast."

"fast,accurate, widely used, well documented"

STAR

"STAR is ridiculously fast, and I have access to sufficient RAM for running it. However I often use Tophat because it is so well established in many pipelines."

"speed."

BWA

"Both are fast (BWA and Bowtie). Bowtie works in colorspace"
"Affidabile"

GSNAP

"I work with maize and there is so much variation (both SNPs and indels) between cultivars. GSNAP seems to do the best job of equal alignment of reads from non-reference lines."

X-mate *"Experiance"*

Shrimp *"works well for color space (solid sequencing)"*

RUM *"Profiling results from the paper describing the rum algorithm."*

NovoAlign

"Fast, commercial support, gapped alignments. Has MPI mode, reads from gzip'd files, does quality trimming for you, correctly determines the 33/64 quality encodings."

Lifescope *"colorspace alignment"*

GEM

"GEM is way better than the others: http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.2221.html"

MapSplice

"Quick and accurate alignments to genome, collaboration with authors of MapSplice"

Segemehl

"- mapping with mismatches and indels - allows errors in seed - returns all multiple mapping loci"

IndexDP

"Works good for my platform (namely Helicos)"

Question 4: What is your preferred read-counting methodology?

	Response Percent	Response Count
Cufflinks/CuffDiff	57.1%	44
HTSeq-count	37.7%	29
easyRNASeq	5.2%	4
[redacted]	0.0%	0
Other (please specify) Show Responses		19
answered question		77
skipped question		18

Other read counters (number of times mentioned): eXpress(2), RSEM(4), bedtools(3), internal tool/ inhouse-scripts(9), Samtools/Rsamtools(2)

Question 5: Reasons for choice of read counter

Cufflinks/CuffDiff

"Cufflinks would be 1st choice but is too much of a black box, hence rather read-based"

"ease of use"

"Easy. Starting to use HTSeq-counts too."

"The Cuff" package provides the most mathematically sound model and bias analysis."

"Ease of use (for me)"

"becuse of tophat"

"is simple to use, there are some option for bias correction"

"Widely used tuxedo package"

" Mostly just because everyone else uses this method"

"Easy workflow"

"Again, experience using the system and ability to identify and quantify novel splice variants"

"It's popular"

"Everyone is familiar with Cufflinks, so my choice rarely needs to be defended (even of other methods give more accurate gene quantification results)."

"Cufflinks goes along with Tophat. It seems to be having problems lately so I may switch."

HTSeq-count

"We are focused on expression at the gene rather than the transcript level, so htseq-count with the union method is the most straight forward of assigning counts to genes."

"I don't believe the output of Cufflinks/diff, plus it is slow and has several issues. I just want simple counts."

" More accurate than CuffLinks/CuffDiff"

"ease of use ; documentation"

"fast and accurate"

"Known"

"HTSeq-count for genes, RSEM / eXpress for isoform level counts. Actually the latter would probably be better for genes as well but I just started using them and HTSeq is in production in our pipeline"

"Historical"

"I used cufflinks/cuffdiff originally, but I've found so many issues with how those tools do things that I no longer trust them. HTSeq-count is fast, easy to use, and performs in a way where it's easy to check for errors."

easyRNASeq

"Is esasy to use and is possible to pipe with edgeR and DESeq"

"Cuffdiff gives weird output, not consistent across versions, seems to change all the time."

eXpress *"Depends on the analysis I am doing (Cufflinks for FPKMs, eXpress for read counts)"*

bedTools *"BEDtools is easy to use and reliable"*

RSEM *"Also Trinity's built-in."*

In-house method/ own scripts

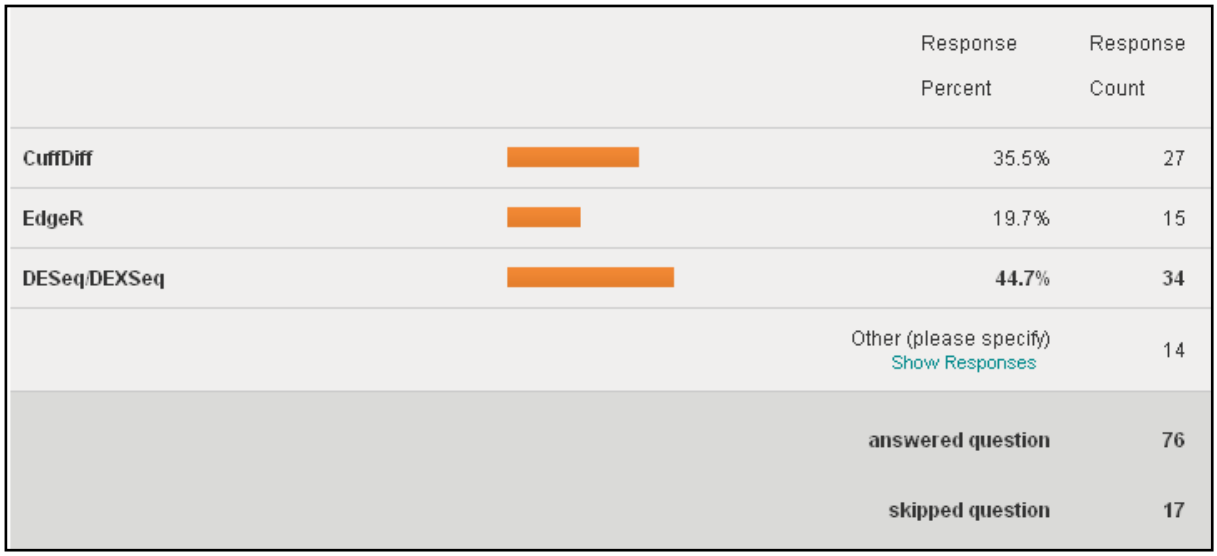
"I am picky about my stats"

"Because programs available at the time were not sufficient"

"Maximum ability to customize to my own needs."

"I'm working with small RNA-seq data and want to normalize for multiple mappings"

Question 6: What is your preferred method to estimate differential expression?



	Response Percent	Response Count
CuffDiff	35.5%	27
EdgeR	19.7%	15
DESeq/DEXSeq	44.7%	34
Other (please specify) Show Responses		14
answered question		76
skipped question		17

Other DE methodologies (number of times mentioned): SAMSeq(1), FPKM -> ANOVA + bonferroni(1), QuasiSeq(1), FluxCapacitor(1), Limma(1), ALDEx(1), Custom/in-house(3), EBSeq(1), all three listed(1)

7. Reasons for choice of differential expression methodology

CuffDiff

"becuse of tophat"

"Comfortable with more informative result"

"Often I can not, unfortunately, fulfill the replicate requirements of DESeq and DEXSeq so my analyses are more exploratory in nature."

"Easy to do it. But i think using EdgeR or DESeq would be more flexible and a better choice."

"To capture expression of novel transcripts and isoforms"

"Can id novel splice variants and then reannotate the gtf thus estimating expression of new transcripts"

"Easy to use with Cufflinks."

EdgeR

"Scripts were built for me."

"For our purposes (32-50 bp reads, 10-20M reads per replicate, single end reads, interested in gene-level expression) cuffdiff gave many false positives and other problems. We chose edgeR over DESeq because DESeq was too conservative."

"well documented"

"I found EdgeR more reliable than the other two for DEG"

"First one made and most informative."

DESeq/DEXSeq

"DESEq accepts custom input"

"Custom/in-house"

"None have adequate methods is dealing with complex designs"

"ease of use."

"Works best in my experience (DESeq on eXpress output)"

"DESeq and EdgeR are more or less equivalent in my book. I simply do not trust the results from cuffdiff. Something I do do on occasion, is use the GTF file generated by cufflinks as an annotation for htseq-count and DESeq. Depending on the organism and annotation, you can pick up unannotated exons (not to mention getting better coverage of UTRs)."

QuasiSeq *"Advice from a colleague who knows more about statistics than I do."*

All three listed methods *"I'm triying to get many point of view"*

No Preference

"they're not doing the same thing: transcript-level for Cufflinks, gene-level for edgeR/DESeq"

"edgeR for factorial designs (DESeq is fine too) SAMSeq when I have many replicates - less problems with outlier values"

Question 8: Which annotation resource do you use?

	Response Percent	Response Count
UCSC	22.4%	13
RefSeq	25.9%	15
Ensembl	46.6%	27
Gencode	5.2%	3
Other (please specify) Show Responses		6
answered question		58
skipped question		35

Others: JGI, SEED, KEGG

Comments

"Not many resources for non-model organisms"

"Combined UCSC, RefSeq and Ensembl"

"refSeq for simple gene expression as it is a simple annotation with not too much complexity; GENCODE if I want to do splicing analysis or something that requires comprehensive annotation on ncRNAs"

Question 9: What software do you use for downstream analyses? e.g. pathway enrichment

	Response Percent	Response Count
GOSeq	68.9%	31
Ingenuity (IPA)	33.3%	15
Genego MetaCore	8.9%	4
Other (please specify) Show Responses		21
answered question		45
skipped question		48

Others (number of times mentioned): DAVID (5), KEGG, Blast2GO(2), GSEA(1), CLC Genomics(1), ToppGene/ToppCluster(1), GOstats(1), TopGO(1), MEGAN(1)

Comments

"Ingenuity is not reliable We use a variety of local pathway analysis algorithms"

"I do more exome-seq analysis, so haven't gotten into some of the downstream tools for RNA-seq"

"Custom because there are not many resources for non-model organisms. Any pipeline should be fully customizable for this reason."

"Depends on the analysis I am doing"

"Blast2Go ... but I am getting dissatisfied with it."