best practices for RNASeq analyses and automated workflow design (e.g., QC, single-cell workflows, normalization, statistics, alternative splicing, downstream automated functional analyses)

a few publications below describing Kallisto and a very similar approach Salmon.

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374695/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.ncbi.nlm.nih.gov_pmc_articles_PMC5374695_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=ROXOHj3MUybwZcieDApWqp-0jkXhPpEFT4jC-8F8TSg&e=)

[https://www.ncbi.nlm.nih.gov/pubmed/27043002](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.ncbi.nlm.nih.gov_pubmed_27043002&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=G0OwGdddPy-AHpx3nc0IDz0rdLd1n7juzyVqTJYtQhY&e=)

[https://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.nature.com_nbt_journal_v32_n5_full_nbt.2862.html&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=JPKPrRbdDsGE6DGAX4Vgq32IfxZjn9gxSgk0B4iTf74&e=)

There are a number of publications that compare various aspects of alignment with STAR versus other approaches. Most of these focus on the ability to identify known and novel spliced junctions, deletions and in some cases fusions, in well annotated versus poorly annotated genomes (lots of scaffolds). STAR typically performs very well for different genomes at junction discovery and finding small deletions. Accurately detecting fusions, circular and novel genes typically can work with all of these BAM file outputs. When running STAR, it is important to include the flag to predict strand orientation for each read (--outSAMstrandField intronMotif) and of course a gtf file to obtain improved alignments to a transcriptome reference.

a few publications below describing Kallisto and a very similar approach Salmon.

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374695/](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.ncbi.nlm.nih.gov_pmc_articles_PMC5374695_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=ROXOHj3MUybwZcieDApWqp-0jkXhPpEFT4jC-8F8TSg&e=)

[https://www.ncbi.nlm.nih.gov/pubmed/27043002](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.ncbi.nlm.nih.gov_pubmed_27043002&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=G0OwGdddPy-AHpx3nc0IDz0rdLd1n7juzyVqTJYtQhY&e=)

[https://www.nature.com/nbt/journal/v32/n5/full/nbt.2862.html](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.nature.com_nbt_journal_v32_n5_full_nbt.2862.html&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=JPKPrRbdDsGE6DGAX4Vgq32IfxZjn9gxSgk0B4iTf74&e=)

There are a number of publications that compare various aspects of alignment with STAR versus other approaches. Most of these focus on the ability to identify known and novel spliced junctions, deletions and in some cases fusions, in well annotated versus poorly annotated genomes (lots of scaffolds). STAR typically performs very well for different genomes at junction discovery and finding small deletions. Accurately detecting fusions, circular and novel genes typically can work with all of these BAM file outputs. When running STAR, it is important to include the flag to predict strand orientation for each read (--outSAMstrandField intronMotif) and of course a gtf file to obtain improved alignments to a transcriptome reference.

Gene-level abundance estimates versus transcript-level

analyses in salmon (fig2), and featurecounts versus

salmon (fig3):

[https://f1000researchdata.s3.amazonaws.com/manuscripts/8143/fa54b4f8-fda8-465b-8db5-b544644b10ee\_7563\_-\_mark\_robinson.pdf?doi=10.12688/f1000research.7563.1](https://urldefense.proofpoint.com/v2/url?u=https-3A__f1000researchdata.s3.amazonaws.com_manuscripts_8143_fa54b4f8-2Dfda8-2D465b-2D8db5-2Db544644b10ee-5F7563-5F-2D-5Fmark-5Frobinson.pdf-3Fdoi-3D10.12688_f1000research.7563.1&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=s584ynXCebb97NA010XctI7Qpm2ZcxY2kPS1Y2IlY3E&e=)

In table 2 for the three types of comparative analyses

discussed in the manuscript (DGE, DTE and DTU) the best

tool is given for each case.

DESeq2 requires raw counts as input and the option

"--quantmode" can be used in STAR to output a table

with raw reads. Since "Gene abundance estimates are

more accurate than transcript abundance estimates"

the STAR-DESeq2 combination still seems good option

when it comes to speed and accuracy in pipelining.

examples:

--quantMode GeneCounts  or

--quantMode TranscriptomeSAM

This combination STAR-DESeq2 works accurately in

producing DE genes lists, but doesn't produce

alternative splicing and isoform information on

a transcript level.

Has Kallisto been compared in alternative splicing

analysis versus some other tools such as PSG or Miso?

I've used the latter with splicing, in combination

with Trinity and RSEM but I'm not sure if these are

really fast enough for the purpose of pipelines for

biominer.

It seems sleuth in combination with Kallisto should

work for transcript-level analyses and these tools

were tested on publicly available data here:

[https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1357-2](https://urldefense.proofpoint.com/v2/url?u=https-3A__bmcbioinformatics.biomedcentral.com_articles_10.1186_s12859-2D016-2D1357-2D2&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=sq39SDUM1o8lMSNb2iA40ry1LepQUQT1IMVyN2pu9O4&e=)

|  |
| --- |
| [The Lair: a resource for exploratory analysis of published ...](https://urldefense.proofpoint.com/v2/url?u=https-3A__bmcbioinformatics.biomedcentral.com_articles_10.1186_s12859-2D016-2D1357-2D2&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=sq39SDUM1o8lMSNb2iA40ry1LepQUQT1IMVyN2pu9O4&e=)  [bmcbioinformatics.biomedcentral.com](http://bmcbioinformatics.biomedcentral.com)  The Sequence Read Archive (SRA) is a public repository for sequencing data that has become an important archival resource for reads associated with published papers. |

With kallisto and sleuth raw RNA-Seq reads can be

converted into a complete analysis in a matter of

minutes like this:

[https://liorpachter.wordpress.com/2015/08/17/a-sleuth-for-rna-seq/](https://urldefense.proofpoint.com/v2/url?u=https-3A__liorpachter.wordpress.com_2015_08_17_a-2Dsleuth-2Dfor-2Drna-2Dseq_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=64shtfrZlYnLmQ7AeqM6qxODVDAYoDtW2mgepmMLN8Y&e=)

|  |  |
| --- | --- |
| [https://liorpachter.files.wordpress.com/2015/08/sleuth.jpg](https://urldefense.proofpoint.com/v2/url?u=https-3A__liorpachter.wordpress.com_2015_08_17_a-2Dsleuth-2Dfor-2Drna-2Dseq_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=64shtfrZlYnLmQ7AeqM6qxODVDAYoDtW2mgepmMLN8Y&e=) | [A sleuth for RNA-Seq](https://urldefense.proofpoint.com/v2/url?u=https-3A__liorpachter.wordpress.com_2015_08_17_a-2Dsleuth-2Dfor-2Drna-2Dseq_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=64shtfrZlYnLmQ7AeqM6qxODVDAYoDtW2mgepmMLN8Y&e=)  [liorpachter.wordpress.com](http://liorpachter.wordpress.com)  [Update July 15, 2016: A preprint describing sleuth is available on BioRxiv] Today my student Harold Pimentel released the beta version of his new RNA-Seq analysis method and software program calle… |

With example datasets, six independent core-workflows

comprising tools such as Tophat–Cufflink–Cuffdiff,

Subread–featureCounts–DESeq2, STAR–RSEM–EBSeq,

Bowtie–eXpress–edgeR, kallisto–sleuth,

HISAT–StringTie–Ballgown, and embeds itself in

Snakemake, which is a modern pipeline management

system:

[https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw143](https://urldefense.proofpoint.com/v2/url?u=https-3A__academic.oup.com_bib_article-2Dlookup_doi_10.1093_bib_bbw143&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=URAM7otnDMIUckyAjxK8PuwnqRlkDRJF5lm444axzLQ&e=)

The scripts as well as the user manual are freely

available at [https://sourceforge.net/projects/hpprna/](https://urldefense.proofpoint.com/v2/url?u=https-3A__sourceforge.net_projects_hpprna_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=rQmS8rF9-Z8SXrH3l8pJdaC8DD4ySuTtykr7MtMThGc&e=)

|  |
| --- |
| [hppRNA download | SourceForge.net](https://urldefense.proofpoint.com/v2/url?u=https-3A__sourceforge.net_projects_hpprna_&d=DwMF-g&c=P0c35rBvlN7D8BNx7kSJTg&r=MzkOo1ixol3gYzkShGRqf-GZ9_swg7-3yHINCJ2olAs&m=XkK3FCh95Il0S2_fpa24_1vGw3WmlAh5UkZCJV422RY&s=rQmS8rF9-Z8SXrH3l8pJdaC8DD4ySuTtykr7MtMThGc&e=)  [sourceforge.net](http://sourceforge.net)  Download hppRNA for free. A Snakemake-based handy parameter-free pipeline for RNA-Seq analysis. hppRNA - A Snakemake-based handy parameter-free pipeline ... |

complementary methods generally all have different strengths (in no particular order):

1.  DEXSeq

<http://bioconductor.org/packages/release/bioc/html/DEXSeq.html>

                Robust and widely used, but doesn't classify splicing into canonical event types, so results not as intuitive

2. rMATS

<http://rnaseq-mats.sourceforge.net/>

                Robust and widely used, and can identify unannotated splicing events

3.  MISO

<http://genes.mit.edu/burgelab/miso/>

                From the Burge lab, been around for a long time.  Annotation dependent.

4. Spanki

<https://github.com/dsturg/Spanki>

RNA processing beyond alternative splicing

5. Cuffdiff JSD (Jensen-Shannon Divergence metric)

<http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/>

                I use this to find birds-eye bulk differences in splicing.

6. Majiq

<https://majiq.biociphers.org/>

                Produces good visualizations

Finally, here are some reviews (Although a little old now):

                Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems.

                Liu R, Loraine AE, Dickerson JA.

                BMC Bioinformatics. 2014 Dec 16;15:364. doi: 10.1186/s12859-014-0364-4.

<http://www.biomedcentral.com/1471-2105/15/364>

                A survey of software for genome-wide discovery of differential splicing in RNA-Seq data.

                Hooper JE.

                Hum Genomics. 2014 Jan 21;8:3. doi: 10.1186/1479-7364-8-3.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3903050/>

                Methods to study splicing from high-throughput RNA sequencing data.

                Alamancos GP, Agirre E, Eyras E.

                Methods Mol Biol. 2014;1126:357-97. doi: 10.1007/978-1-62703-980-2\_26.

<https://www.ncbi.nlm.nih.gov/pubmed/24549677>

Systematic evaluation of spliced alignment programs for RNA-seq data.

Nat Methods. 2013 Dec;10(12):1185-91. doi: 10.1038/nmeth.2722. Epub 2013 Nov 3.

Engström PG1, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P; RGASP Consortium.

<https://www.ncbi.nlm.nih.gov/pubmed/?term=Systematic+evaluation+of+spliced+alignment+programs+for+RNA-seq+data>