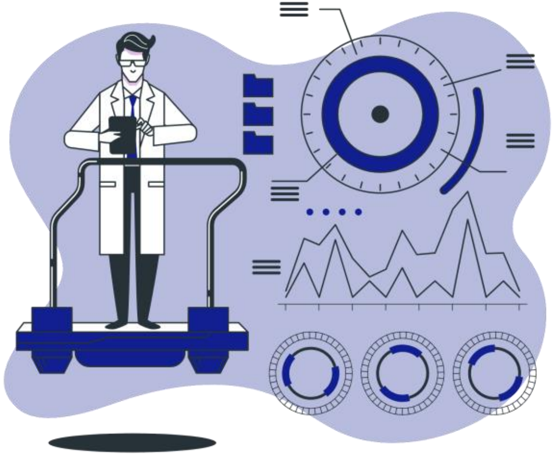
Comparing RNA-Seq Alignment Tools Using a Mouse Breast Cancer Model

Anna Rees | MMG3320 | Spring 2024

Introduction - Background

- Bioinformatic pipelines = time and memory
- Sequence aligners are intended to make the bioinformatic pipeline efficient, user-friendly, and replicable
- **OVERALL GOAL:** compare HISAT2 and STAR alignment tools
    - Same dataset,
    - Same pipeline
- **TESTING**: differences in RNA-Seq analysis outcomes on the same dataset between HISAT and STAR



How do the HISAT2 and STAR aligner tools perform on the same data?

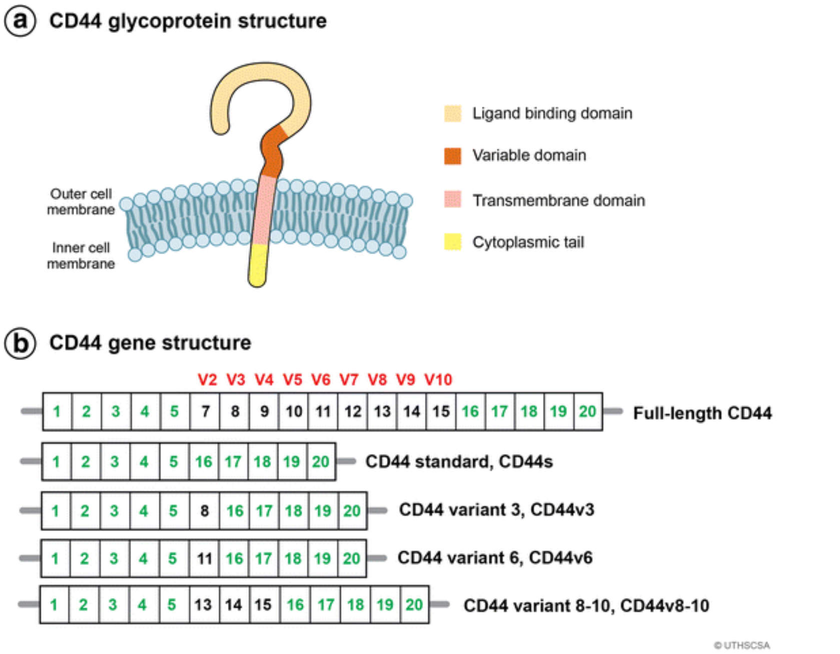
Introduction - Background

Dataset: a transcriptome analysis of metastatic breast cancer in a mouse model

Cancer metastasis: the spread of cancer cells from the source tumor to other tissues

- Known biomarker: CD44
- Sought to ID other cellular biomarkers associated with metastasis

What are the transcriptomic changes that occur during breast cancer metastasis?


a CD44 glycoprotein structure
Outer cell membrane
Inner cell membrane
Ligand binding domain
Variable domain
Transmembrane domain
Cytoplasmic tail
b CD44 gene structure
V2 V3 V4 V5 V6 V7 V8 V9 V10
1 2 3 4 5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Full-length CD44
1 2 3 4 5 16 17 18 19 20 CD44 standard, CD44s
1 2 3 4 5 8 16 17 18 19 20 CD44 variant 3, CD44v3
1 2 3 4 5 11 16 17 18 19 20 CD44 variant 6, CD44v6
1 2 3 4 5 13 14 15 16 17 18 19 20 CD44 variant 8-10, CD44v8-10
© UTHSCSA

CD44 protein and gene structure. Image source: Chen, 2018

Experimental Design

- Sample source: 10 MMTV-PyMT female mice
  - Mice are bred to develop palpable mammary tumors for breast cancer research

The sample extraction method:

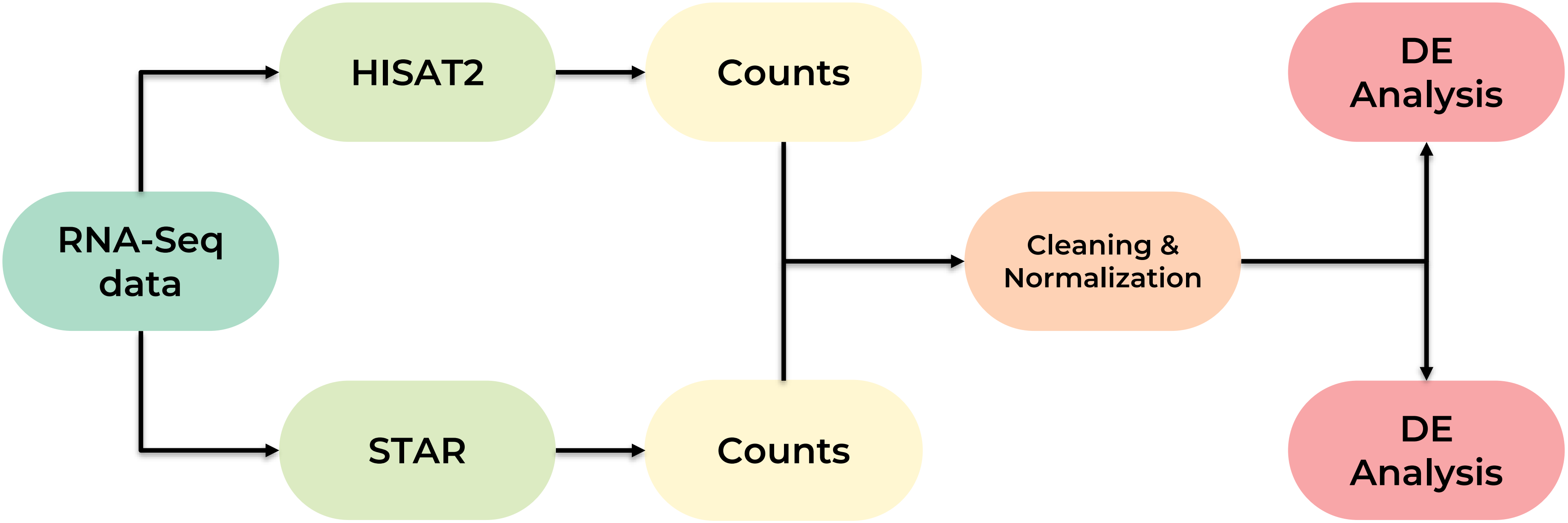
1. harvested >
2. isolated by tissue type >
3. incubated for clonal isolates >
4. RNA extracted >
5. bulk RNA-Seq > scRNA-Seq

8 samples from metastatic tumor-bearing mice derived from Bone Marrow and Lymph Node tissue

GEO Accession code: GSE165393

- Used 8 of the 17 original samples
- Sample conditions:
  - Bone Marrow: Low CD44 level
  - Bone Marrow: High CD44 level
  - Lymph Node: Low CD44 level
  - Lymph Node: High CD44 level
- Paired end reads
- 2 replicates per sample

Bioinformatics Pipeline



Major Findings - Workflow

HISAT2	STAR
✓ Uses known genome annotations to ID splice junctions ✓ Takes less computational energy ✓ Fast x Less customizable x Less sensitive to lower-quality datasets	✓ Uses *de novo* splice-aware aligner ✓ Better equipped to handle low quality datasets ✓ Known for its accuracy ✓ More customizable X Higher memory requirement X Steeper learning curve
• Splice-aware aligners (unlike alignment tools like TopHat) • Most used alignment tools available	

Preliminary Findings

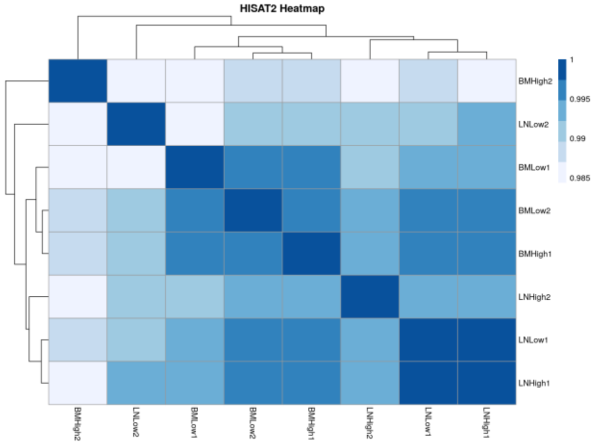
- <u>High variation in the data</u>

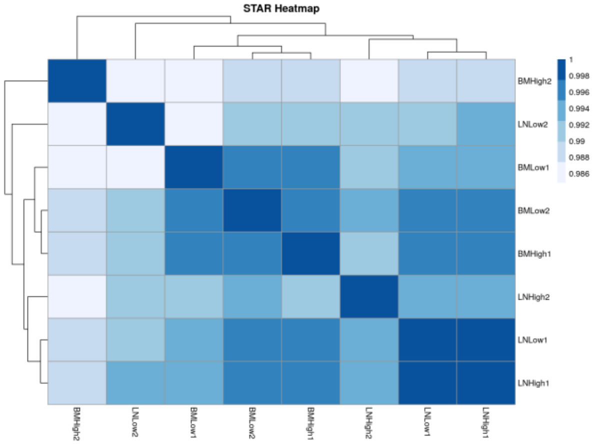
Heatmap:

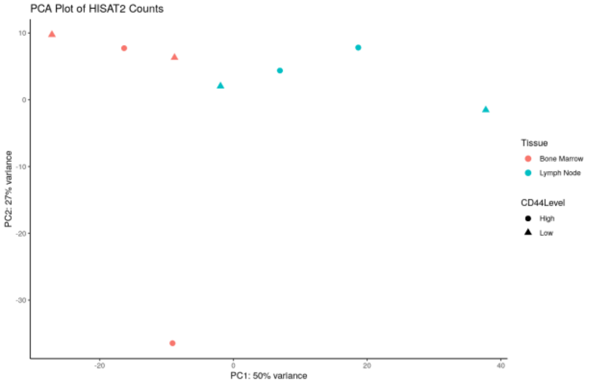
- Higher correlation between LNHigh1 and LNLow1
- HISAT2 and STAR heatmaps draw similar correlations
- High correlations may be due to noise and variability

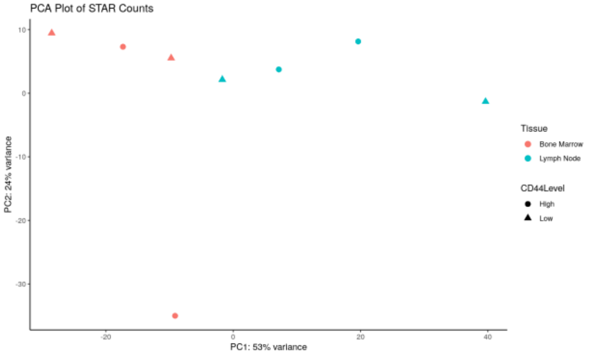
PCA:

- Highly variable (spread out)
- Clustering somewhat by tissue type, not CD44 level

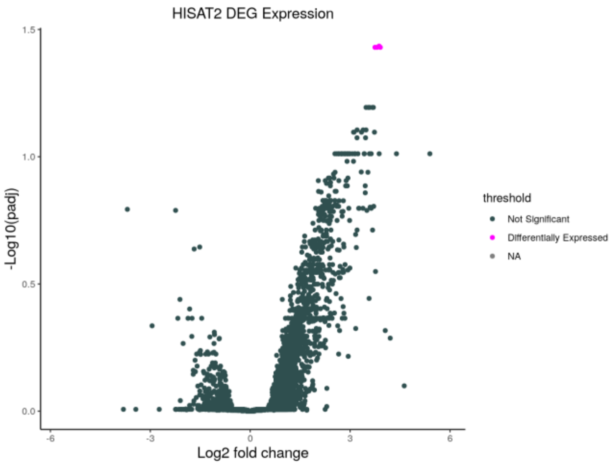

HISAT2 Heatmap

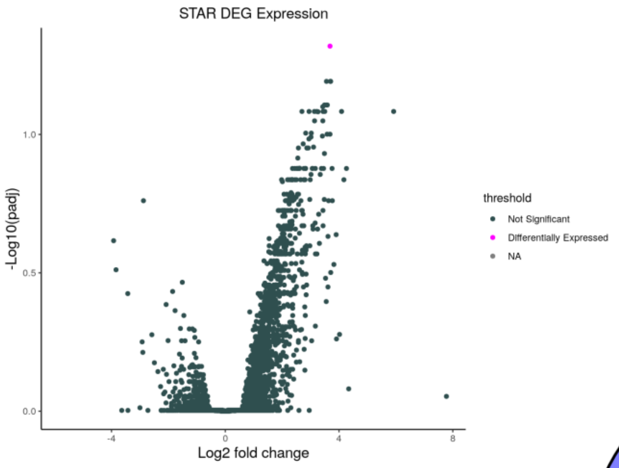

STAR Heatmap


PCA Plot of HISAT2 Counts


PCA Plot of STAR Counts

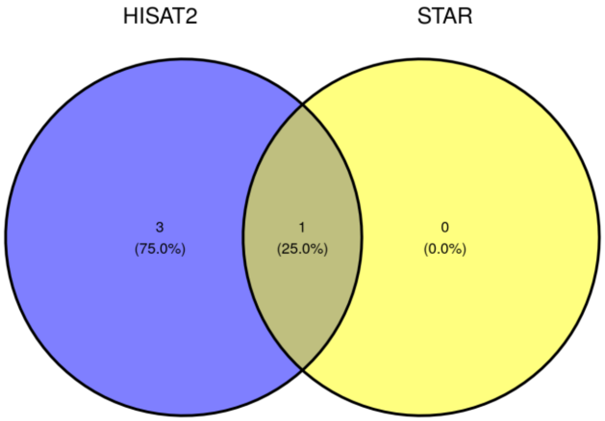
Preliminary Findings



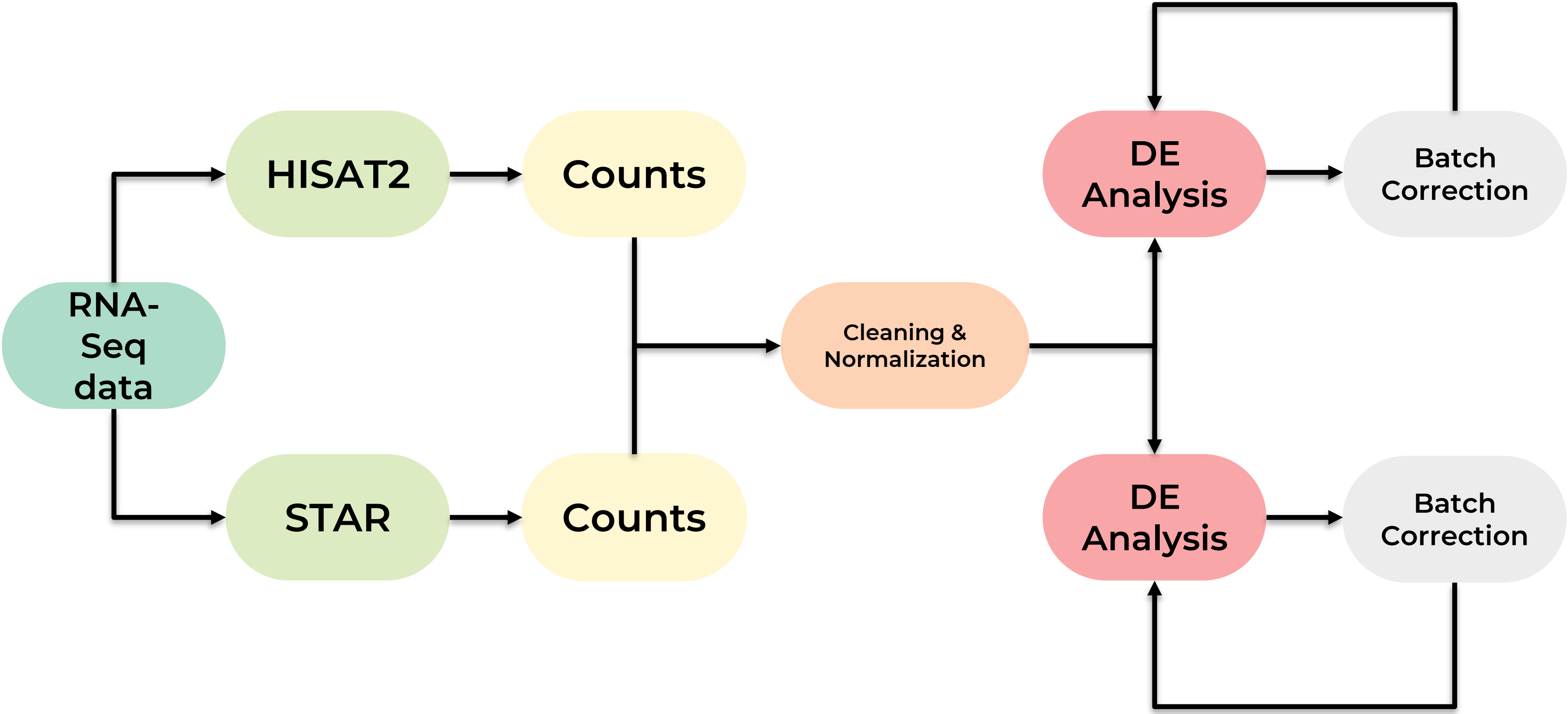


This data is having a batch effect which is impacting the results!

- Threshold for a DEG:
  - Padj < 0.05
  - L2FC > 1
- 1 DEG in STAR
- 4 DEGs in HISAT
- Results likely due to high variation in the data => BATCH CORRECTION



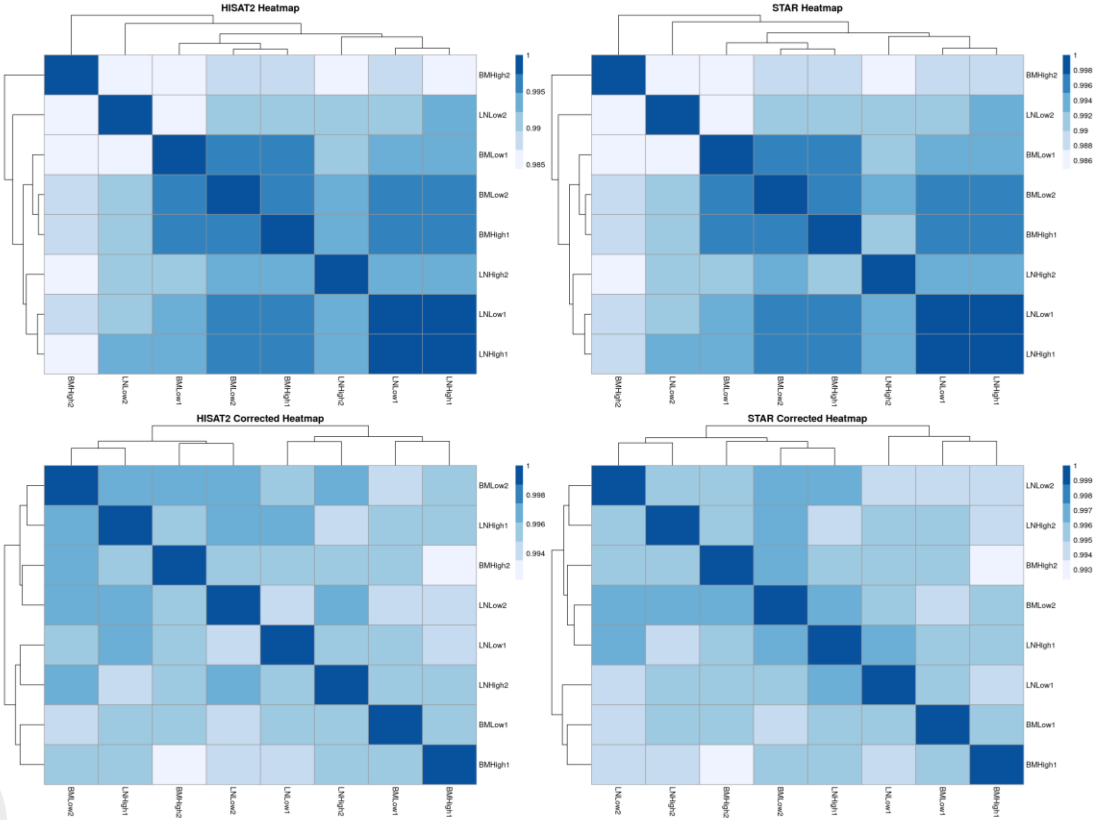
Bioinformatics Pipeline



Major Findings

- After BC: reduced variation in data
  - Less affected by unwanted noise!
- Correlations between different samples are weaker after BC, but more uniform (what we expect to see)

The Heatmap is a sanity check that the batch correction worked

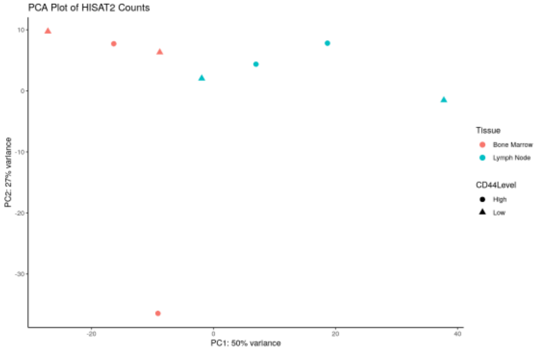


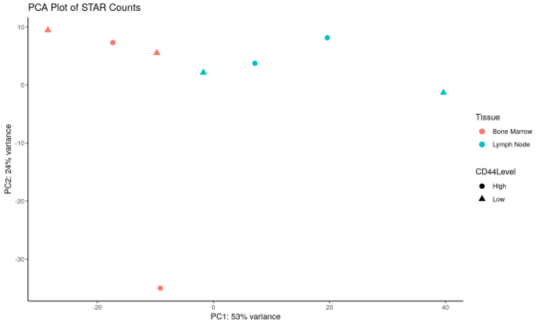
Major Findings

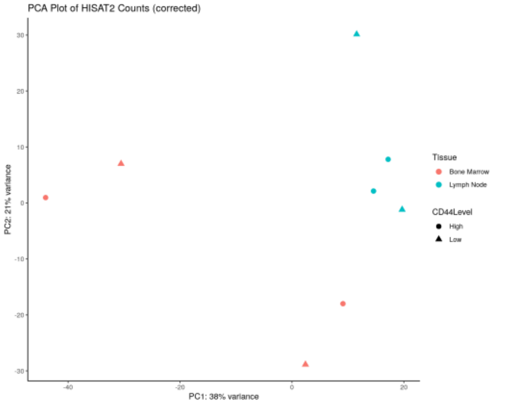
- Clustering between tissue types
    - Specifically Lymph Node
- Still highly variable data
- No clustering by CD44 Level

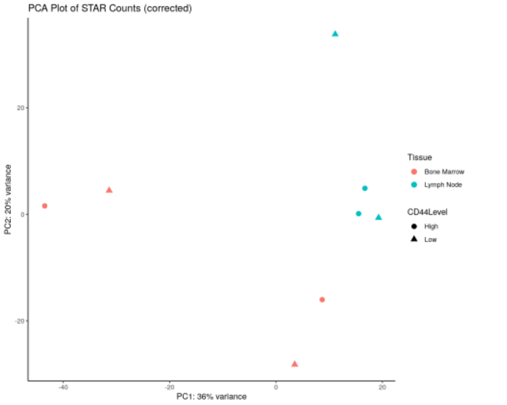
How different are the results of a DE analysis between HISAT2 and STAR data?

Batch Corrected PCA shows clustering in Lymph Node tissue


PCA Plot of HISAT2 Counts

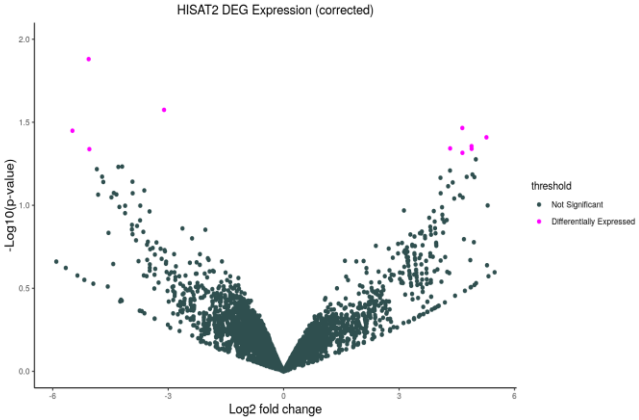

PCA Plot of STAR Counts

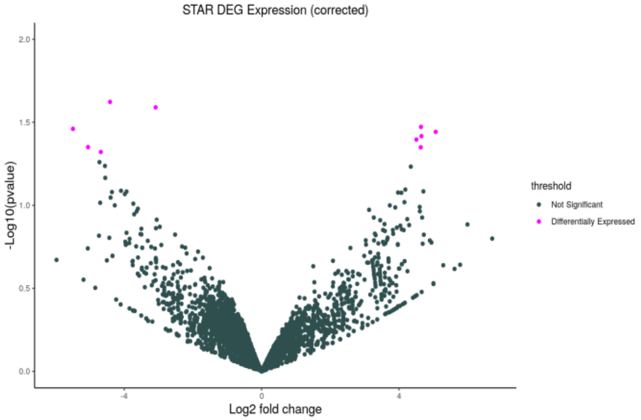

PCA Plot of HISAT2 Counts (corrected)


PCA Plot of STAR Counts (corrected)

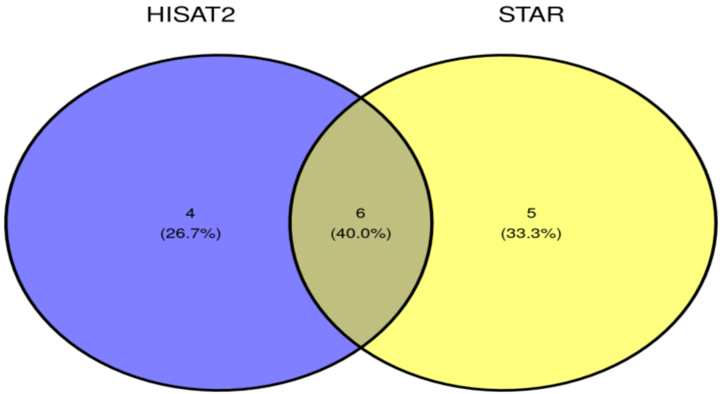
Major Findings

- HISAT2: 10 DEGs
- STAR: 11 DEGS
- Only 6 of the 15 total DEGs agreed between aligners
    - Threshold: p-value < 0.05 and L2FC > 1



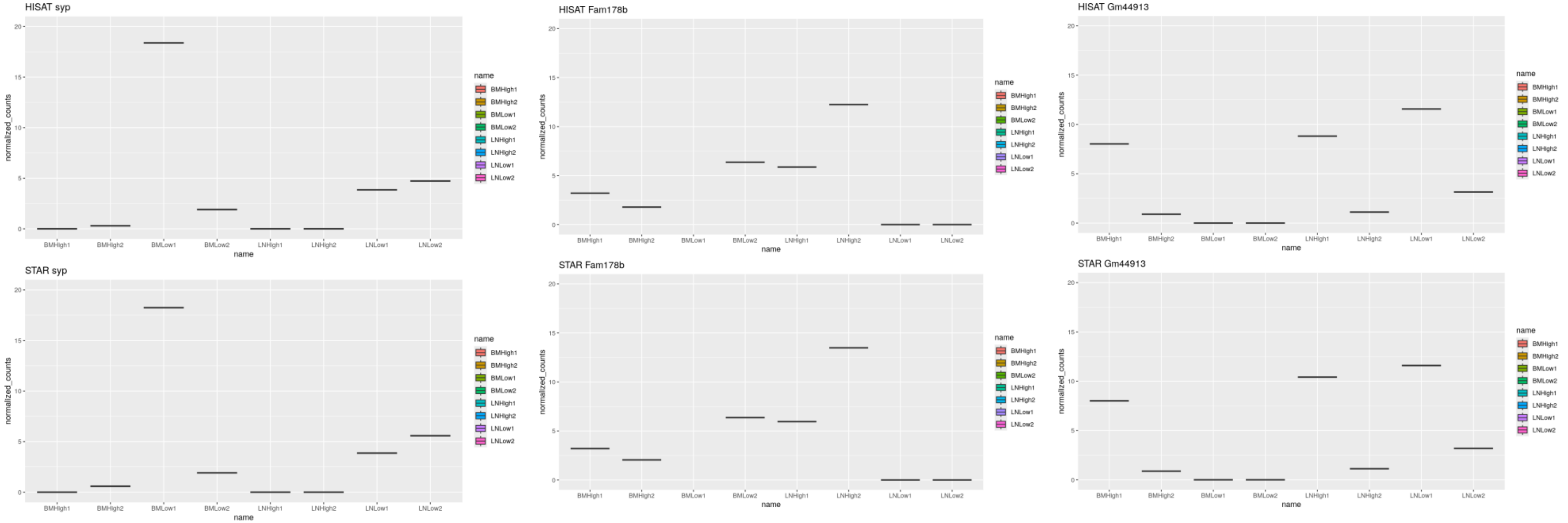


What do the counts for each DEG look like?



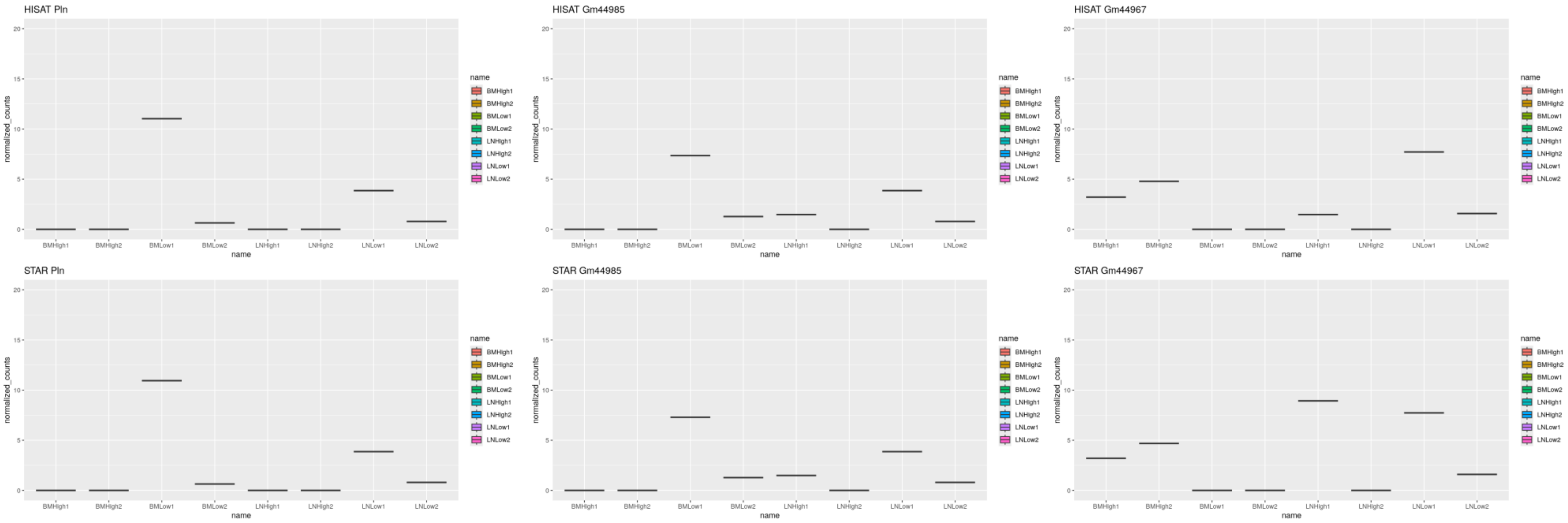
Despite the Batch correction, there are still major differences in results

Major Findings – Agreed DEGs Boxplots



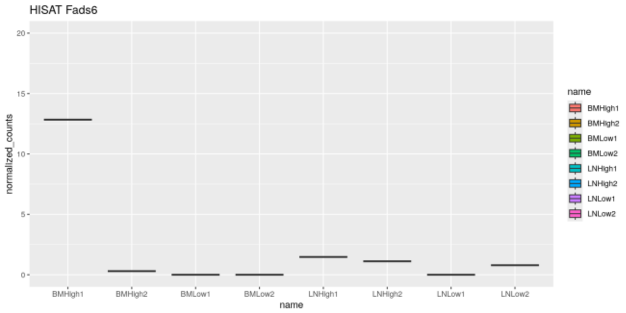
When the aligners agree, so do their normalized counts (for the most part)

Major Findings - Agreed DEGs Boxplots cont.

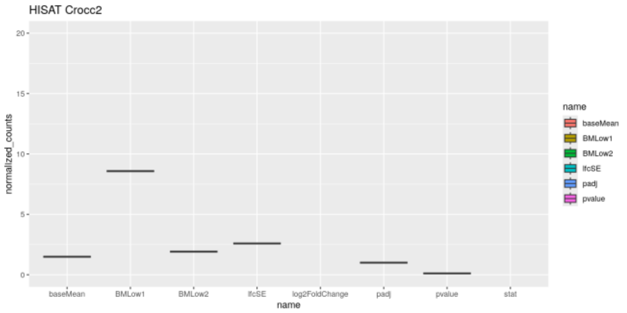


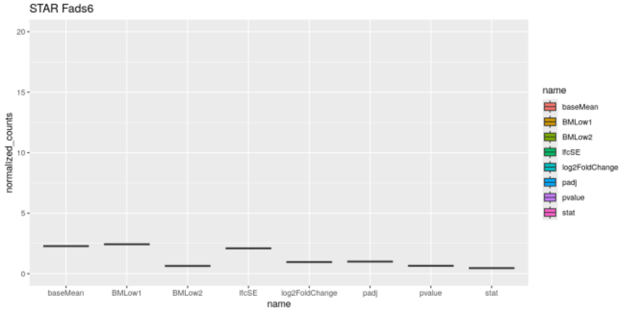
What do the counts for the DEGs the aligners **disagree with** look like?

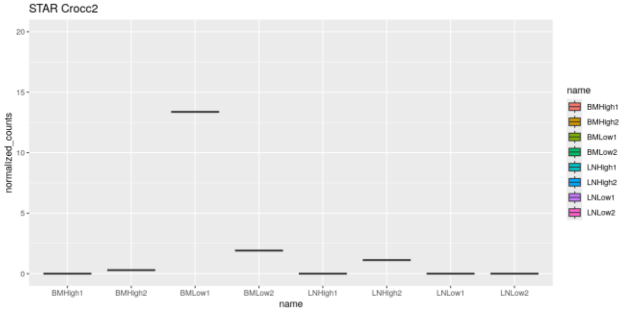
Major Findings – Disagreed DEGs Boxplots



“DEG”







“DEG”

When the aligners disagree, so do their normalized counts

Summary

- The results of DE Analysis can be vastly different depending on the aligner used
- The DEGs identified were 40% agreed, 60% disagreed
  - If agreed, normalized counts agreed
  - If disagreed, normalized counts disagreed
- This conclusion agrees with the original study
- Future research:
  - What is causing the differences in alignment outcomes, and how can they be corrected for the most accurate analysis?
  - What are the implications of the discrepancies in DE analysis results on bioinformatic research moving forward?

Conclusions

- It is up to you (the bioinformatician) to decide what aligner is best for your data
- Batch correction is an important step if necessary
- What to consider when deciding an aligner:
  - Is computational efficiency important to you?
  - How well do you understand these aligners' parameters?
  - What is the quality of your dataset?

Aligners are NOT one-size-fits-all!

References

Bianchi, A., Di Marco, A., & Pellegrini, C. (2023). Comparing Hisat and star-based pipelines for RNA-seq data analysis: A real experience. *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. https://doi.org/10.1109/cbms58004.2023.00220

Chen, C., Zhao, S., Karnad, A., & Freeman, J. W. (2018). The biology and role of CD44 in cancer progression: Therapeutic implications. Journal of Hematology &amp; Oncology, 11(1). https://doi.org/10.1186/s13045-018-0605-5

Ionkina AA, Balderrama-Gutierrez G, Ibanez KJ, Phan SHD et al. Transcriptome analysis of heterogeneity in mouse model of metastatic breast cancer. Breast Cancer Res 2021 Sep 27;23(1):93. PMID: 34579762

Ram, & Lj. (n.d.). Removing batch effects using combat and SVA. Bioinformatics Answers. https://www.biostars.org/p/196430/

Raplee, I. D., Evsikov, A. V., & Marín de Evsikova, C. (2019). Aligning the aligners: Comparison of RNA sequencing data alignment and gene expression quantification tools for Clinical Breast Cancer Research. Journal of Personalized Medicine, 9(2), 18. https://doi.org/10.3390/jpm9020018

The SVA package for removing batch effects and other ... (n.d.). https://bioconductor.org/packages/release/bioc/vignettes/sva/inst/doc/sva.pdf