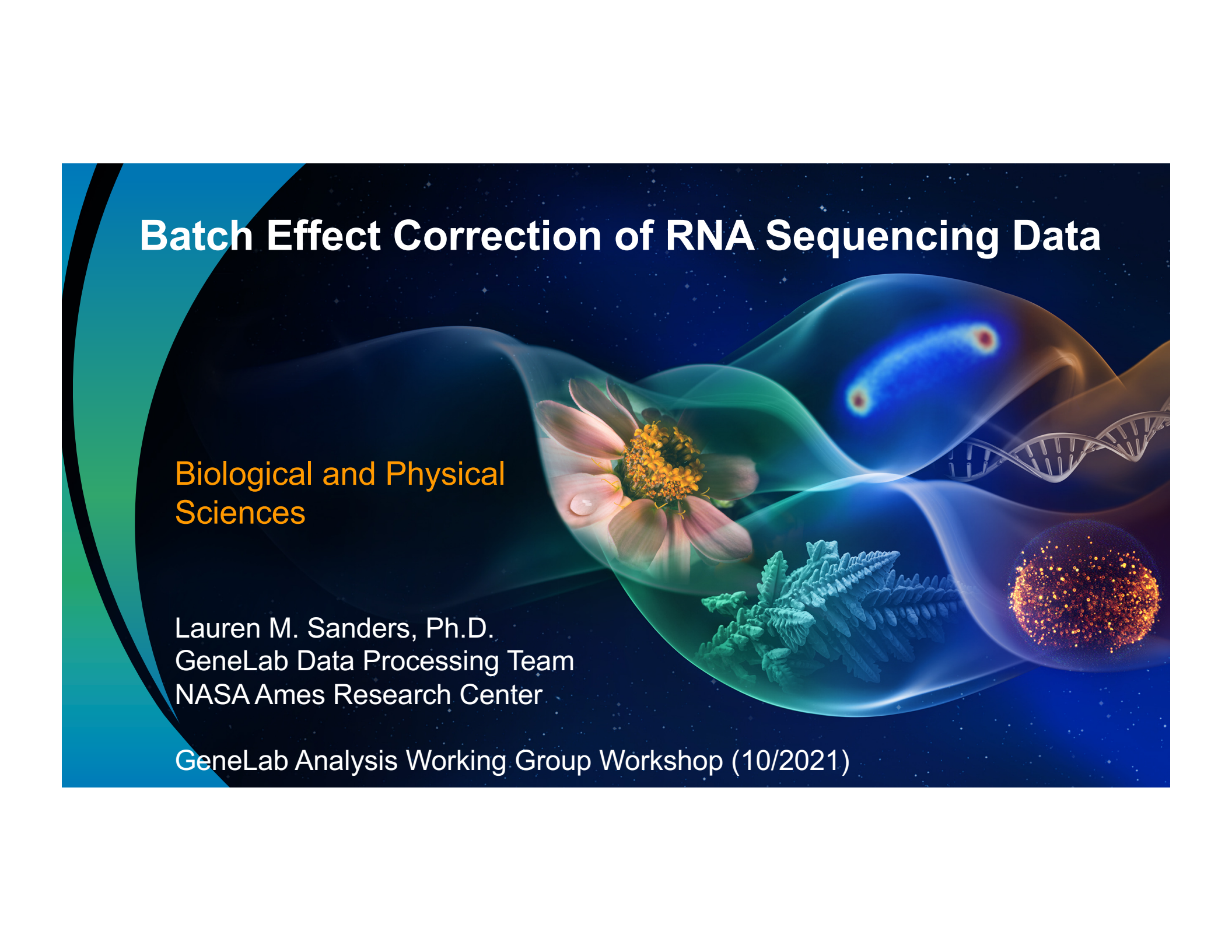


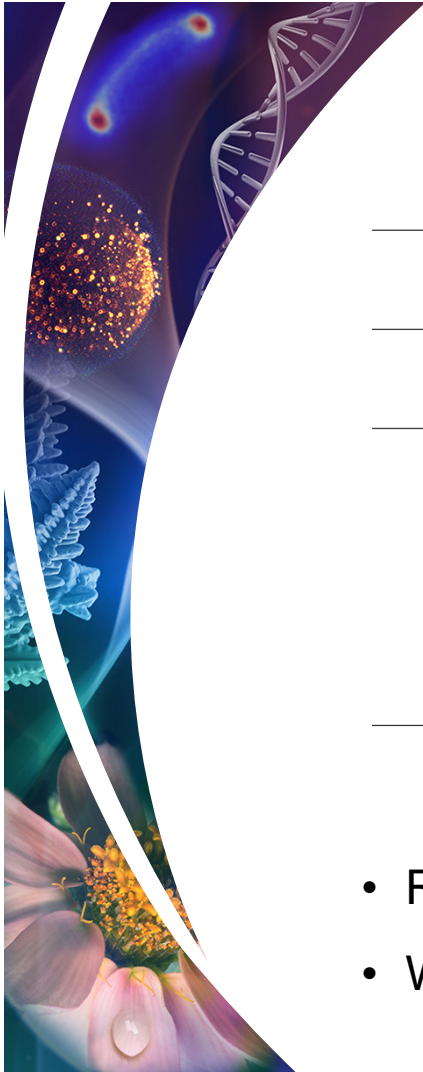
Batch Effect Correction of RNA Sequencing Data

The background of the slide is a deep blue space-like field with white stars. A large, flowing, translucent blue ribbon-like shape curves across the center. Within this shape, there are several elements: a pink flower with yellow stamens, a green fern frond, a white DNA double helix, and a glowing orange sphere with many small dots inside. The overall aesthetic is scientific and futuristic.

Biological and Physical
Sciences

Lauren M. Sanders, Ph.D.
GeneLab Data Processing Team
NASA Ames Research Center

GeneLab Analysis Working Group Workshop (10/2021)



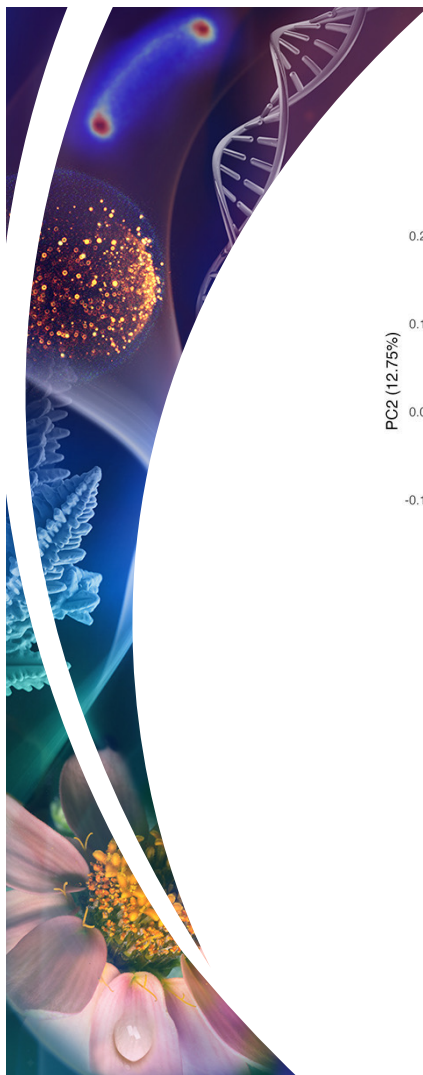
Methods and Variables Evaluated

<i>Tool</i>	<i>Method</i>
Combat	Empirical Bayes
Combat-seq	Negative Binomial
	Empirical Bayes
MBatch	ANOVA
	Median Polish

Corrected variables:

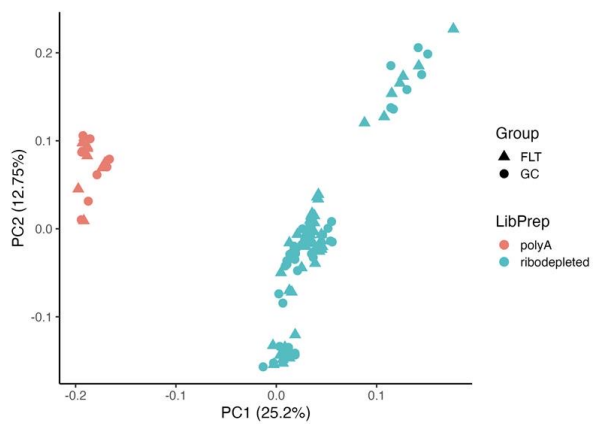
- Library preparation method
- Mission
- ~~Sequencing facility~~

- Result: 10 method / batch variable combinations
- Which is best?

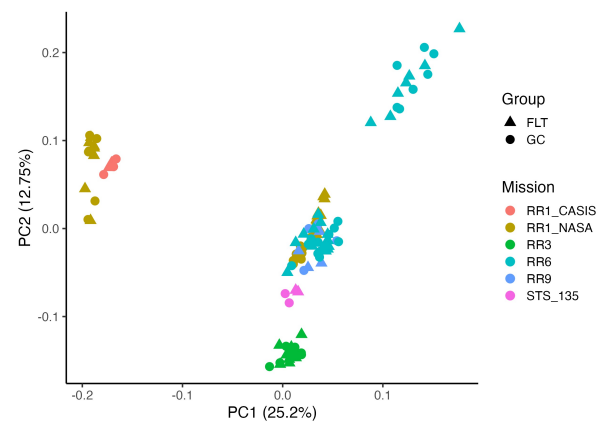


Uncorrected PCA

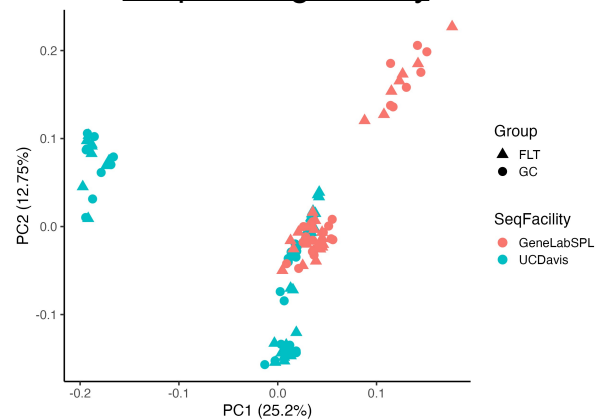
Library Preparation



Mission

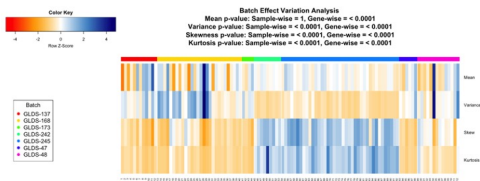


Sequencing Facility

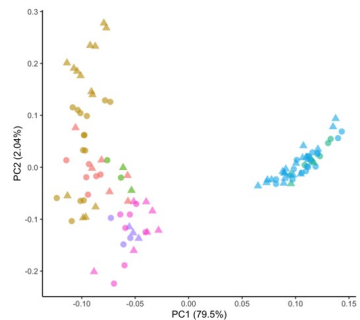


Evaluation Criteria

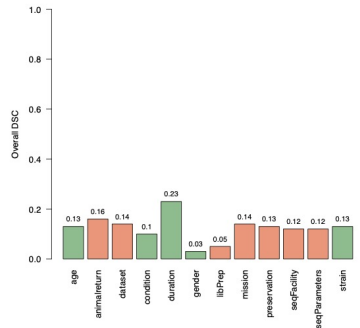
BatchQC



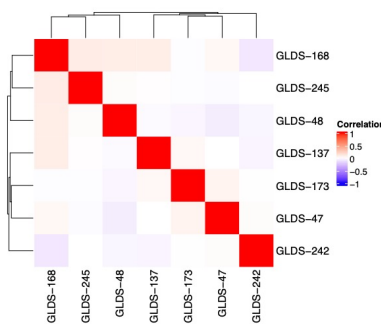
PCA



Dispersion Separability Criterion (DSC)



LFC Correlation

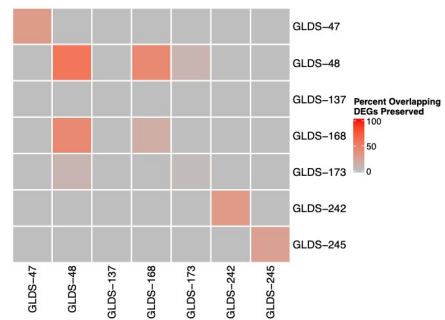


Differential Gene Expression Analysis

Within dataset comparison

	GLDS-47	GLDS-48	GLDS-137	GLDS-168	GLDS-173	GLDS-242	GLDS-245
Uncorrected	14	50	3	1248	520	10	29
ComBat	1 (150)	16 (267)	0 (162)	297 (645)	10 (67)	2 (315)	14 (1584)
ComBat_seq	2 (173)	13 (325)	1 (165)	151 (1027)	2 (88)	0 (2117)	7 (1140)
MBatch_EB	0 (0)	18 (124)	1 (9)	359 (1732)	0 (0)	0 (180)	14 (1372)

Across dataset comparison

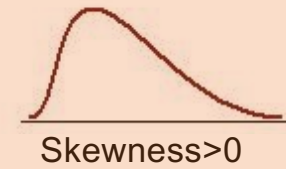
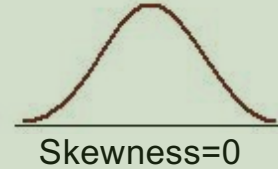


BatchQC

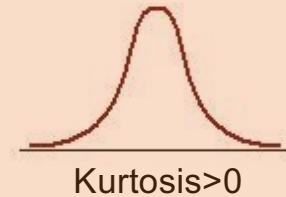
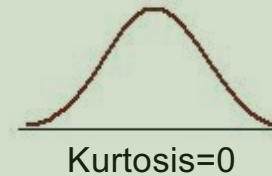
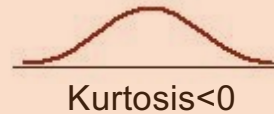
Skew



Normal distribution



Kurtosis

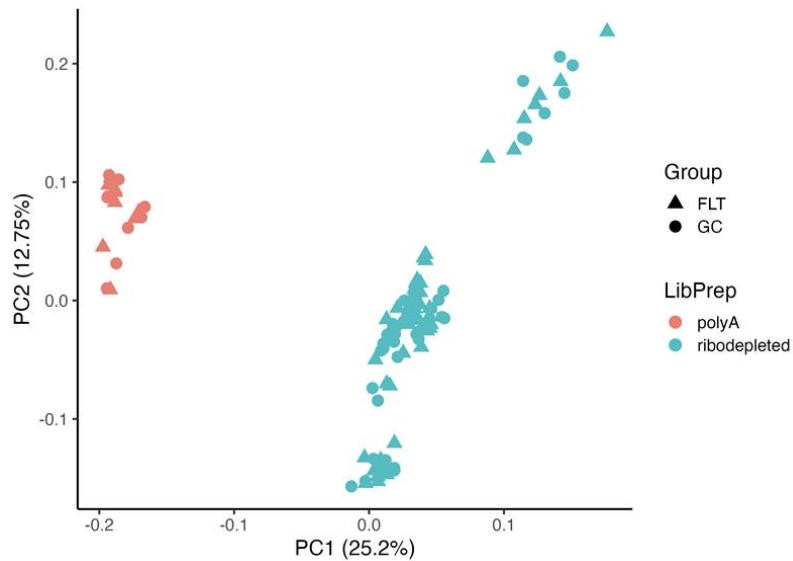


- Assesses how much each sample deviates from the typical sample in a dataset (deviant samples may be affected by batch effects)
- Ideally, we want samples from different technical batches to have similar skew and kurtosis values after correction

For each method/variable combination:

How much closer to “ideal” do samples from different batches get after correction?

PCA

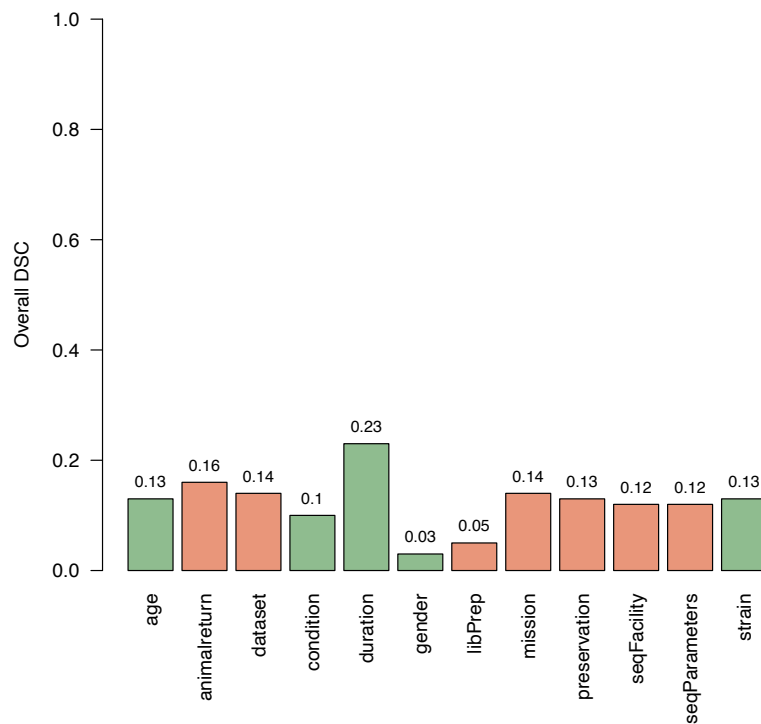


For each method/batch combination:

How much closer does each sample get to samples outside of its technical batch?

How much farther away does each sample get to samples outside of its biological condition?

Dispersion Separability Criterion (DSC)

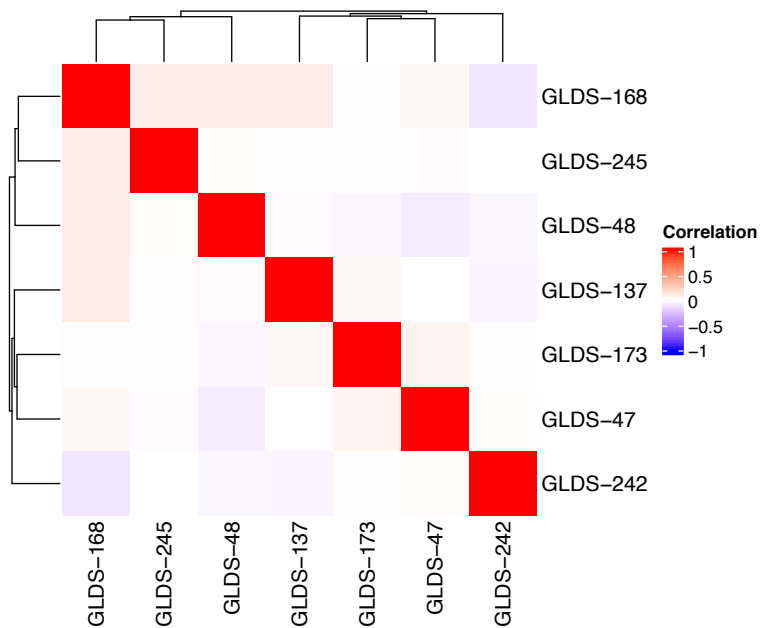


For each method/batch combination:

How much smaller are the technical DSCs after correction?

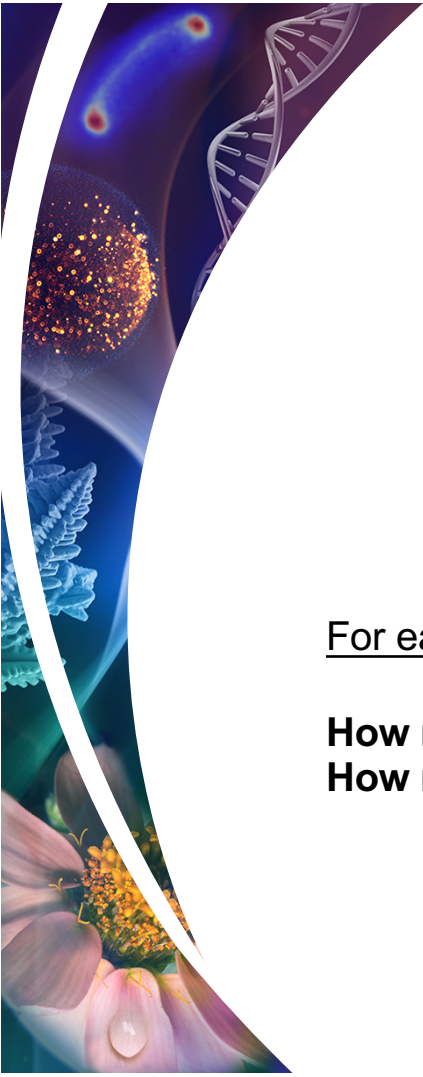
How much greater is the biological condition DSC after correction?

LFC Correlation



For each method/batch combination:

What is the average improvement in LFC for pairwise datasets which come from different batches?



DGE Within-Dataset

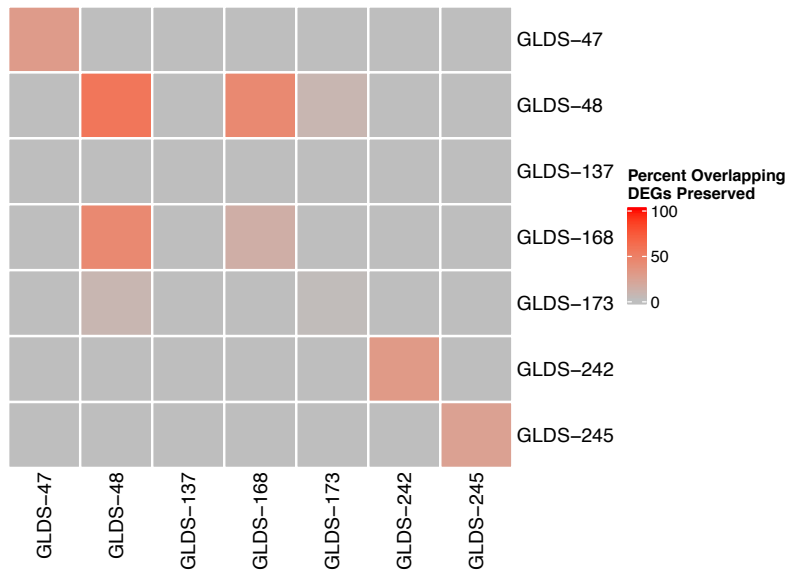
	GLDS-47	GLDS-48	GLDS-137	GLDS-168	GLDS-173	GLDS-242	GLDS-245
Uncorrected	14	50	3	1248	520	10	29
ComBat	1 (150)	16 (267)	0 (162)	297 (645)	10 (67)	2 (315)	14 (1584)
ComBat_seq	2 (173)	13 (325)	1 (165)	151 (1027)	2 (88)	0 (2117)	7 (1140)
MBatch_EB	0 (0)	18 (124)	1 (9)	359 (1732)	0 (0)	0 (180)	14 (1372)

For each method/batch combination:

How many average matching DEGs are there after correction? (maximize)

How many average extra DEGs are there after correction? (minimize)

DGE Across-Dataset

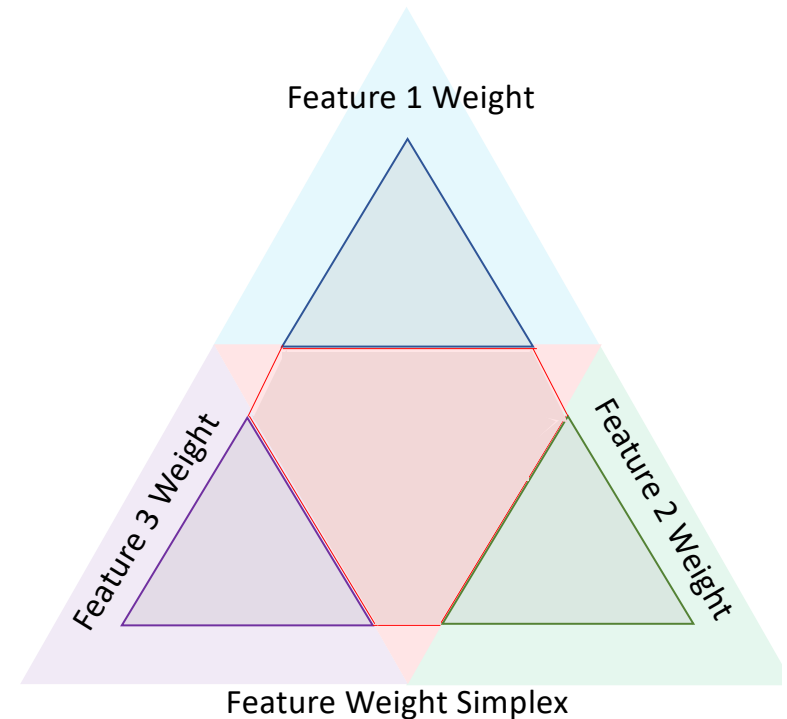


For each method/batch combination:

What is the average improvement in preserved DEGs for pairwise datasets that come from different batches?

Categorization Based Ranking Scheme

- Each method-batch variable pair defines a candidate method, yielding 10 ($=2 \times 5$) candidates
- Each candidate method is assessed in terms of 6 performance measures where each, in turn, may be represented in terms of a set of summary features
- Candidate methods may be viewed as 10 points in some finite (feature) vector space
- Due to partial vector space order, there may not be a unique best performing method
- Any choice of scoring function over the feature space imposes a total order (ranking)
- **Problem:** Which scoring function over the feature space to choose?
- **Solution:** Categorize the space of all possible scoring functions and refer to each inferred category as a preference set (one can also compute the volume of each category)
- Ranking requires the specification of a scoring preference category



- Each subset in the simplex partition denotes a scoring function category
- Functions within the same category are all equivalent as they favor the same method