A standard RNA sequencing assay for space biology

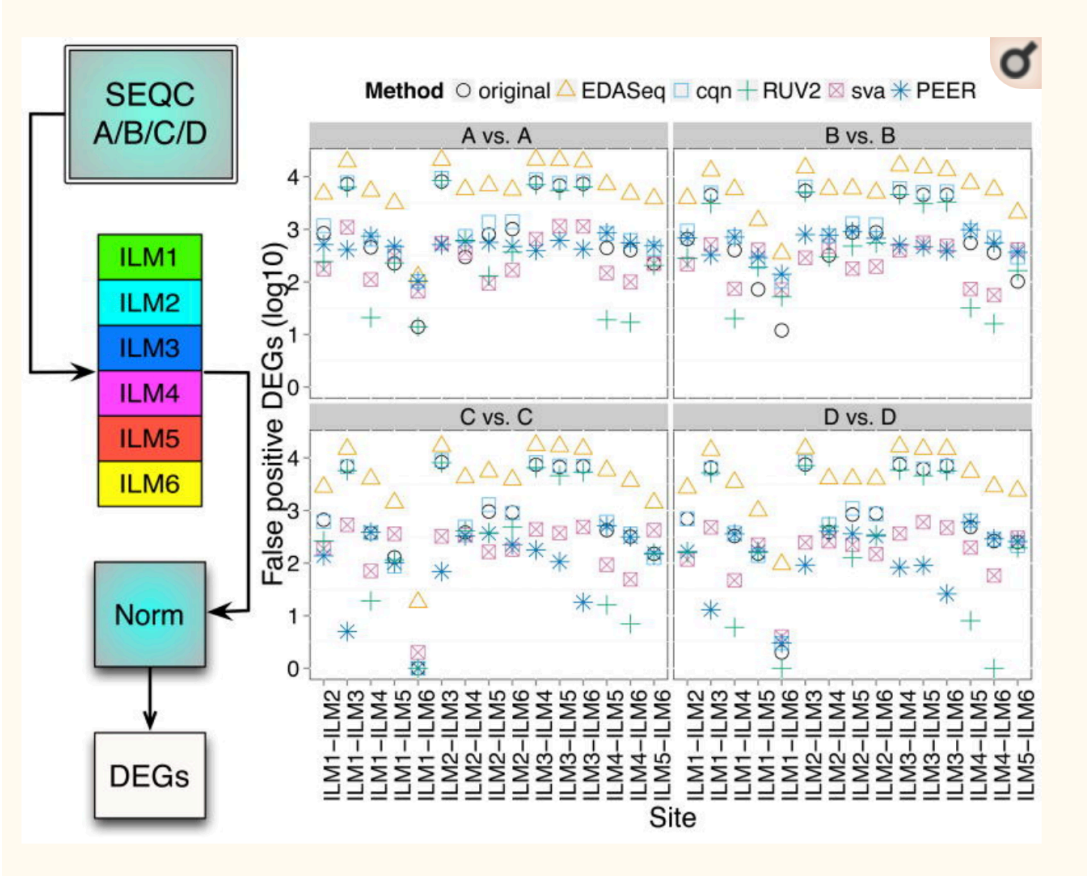
JONATHAN GALAZKA

GENELAB PROJECT SCIENTIST

NASA AMES RESEARCH CENTER

SEQC consortium showed that RNA-seq data is effected by operator identity

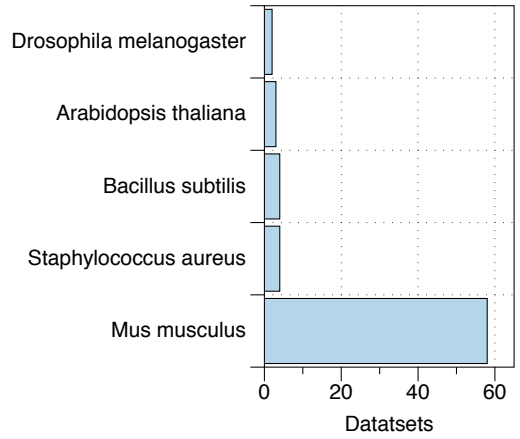
- Same sample sequenced in 6 facilities
- Showed a false discovery rate of 8-20%

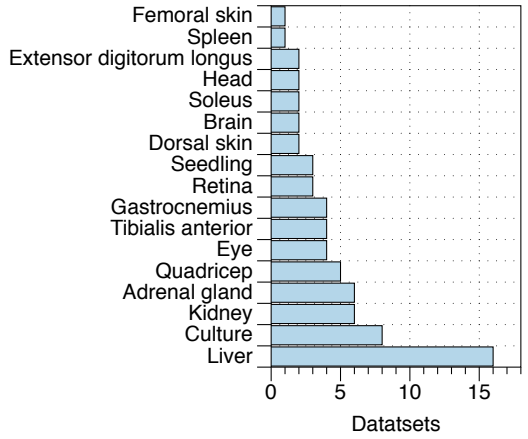


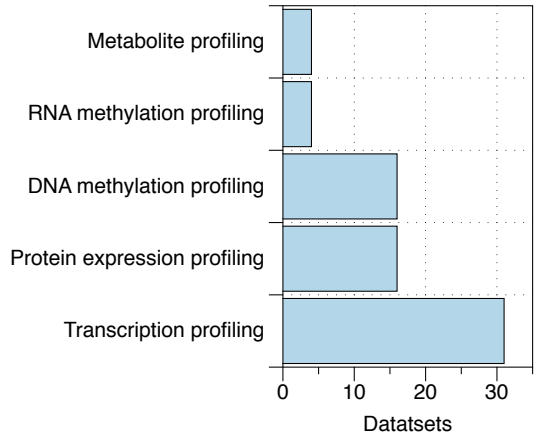
Li S, Łabaj PP, Zumbo P, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol*. 2014;32(9):888–895. doi:10.1038/nbt.3000

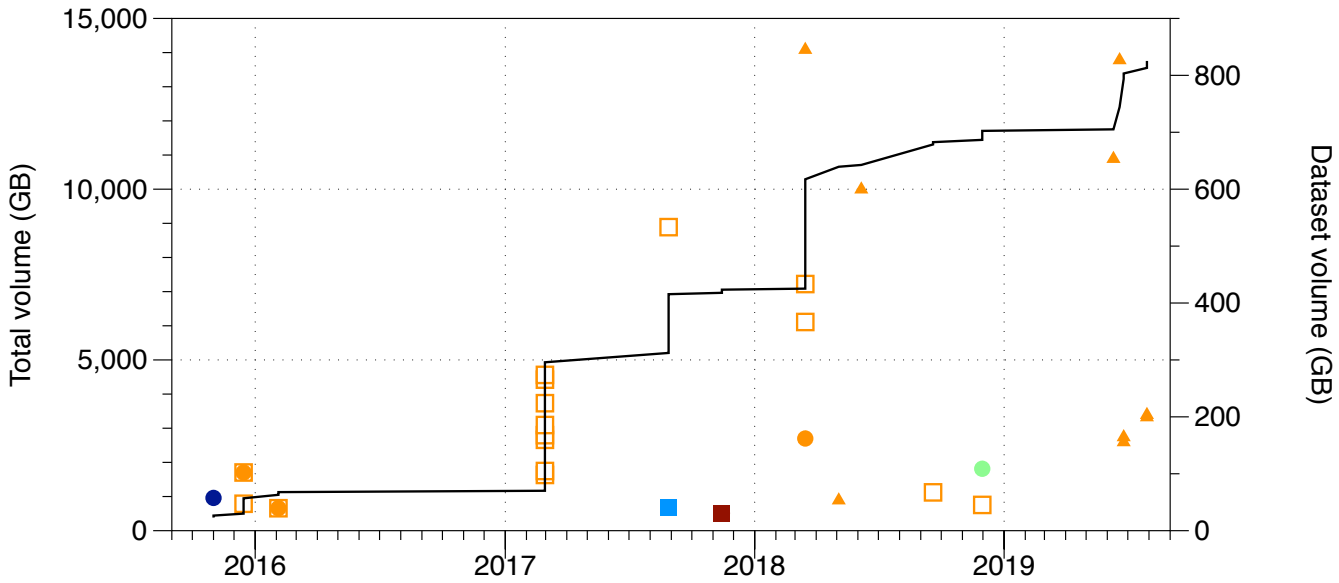
GeneLab Data Generation Summary

Since 2015, GeneLab has helped generate 70 published datasets equaling ~14 TB of data





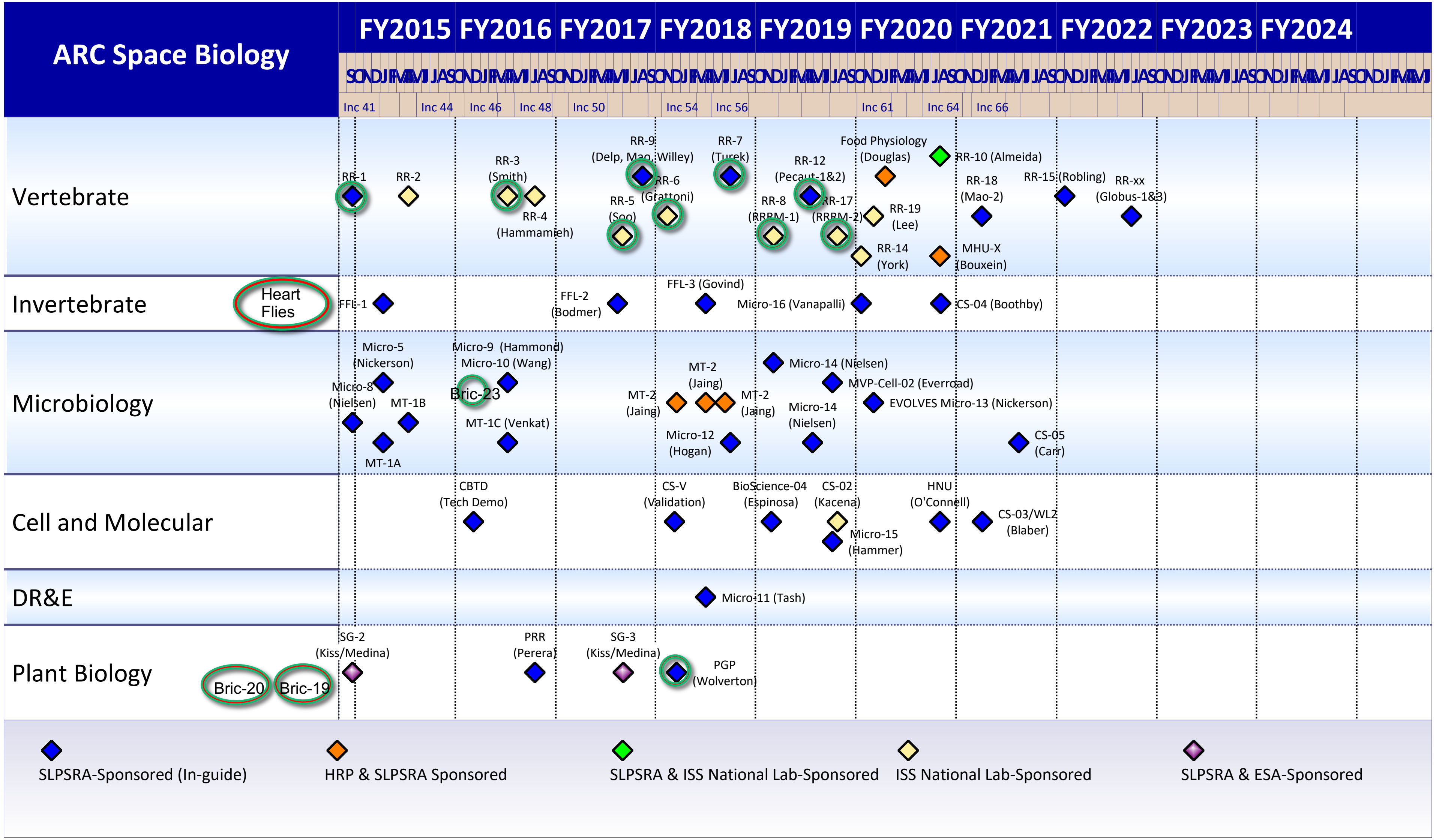




Shapes: Assay type
Triangle: Transcriptional profiling
Open square: DNA methylation profiling
Closed square: Metabolite profiling
Circle: Protein expression profiling
Diamond: RNA methylation profiling

Color: Species
Dark blue: *Arabidopsis thaliana*
Light blue: *Bacillus subtilis*
Green: *Drosophila melanogaster*
Orange: *Mus musculus*
Red: *Staphylococcus aureus*

Ames Space Biology "Fly-off Schedule" 2015 – 2022



= GeneLab data generation (actual and planned)

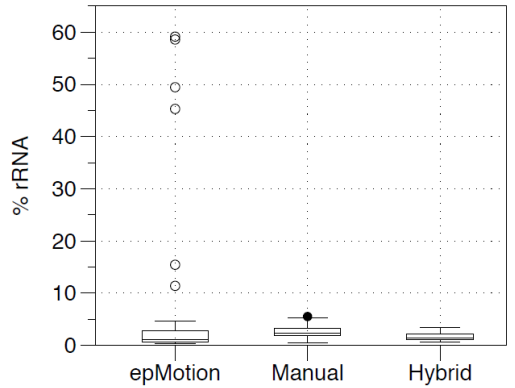
GeneLab Sample Processing Experience: Rodent Tissues

Tissue	Preservation method	Homogenization Method	RNA Extraction Method	Expected RIN
(bact) B. subtilis	Slow freeze in media	Bead Homogenizer	RiboPure RNA Purification Kit	8.1
(bact.)S. aureus	Slow freeze in media	Bead Homogenizer	RiboPure RNA Purification Kit	7.5
Abdominal fat	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.0
Adrenal glands (2)	RNAlater	Polytron	AllPrep DNA/RNA Mini kit	9.5
Adrenal glands (2)	LN2	Polytron	AllPrep DNA/RNA Mini kit	8.6
Brain	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.7
Brown fat	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	8.6
Colon	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	8.1
Extensor digitorum longus (2)	RNAlater	Polytron	AllPrep DNA/RNA Mini kit	9.9
Eye	RNAlater	Polytron	AllPrep DNA/RNA Mini kit	8.0
Eye	LN2	Polytron	AllPrep DNA/RNA Mini kit	7.7
Gastrocnemius	LN2	Polytron	Trizol (no DNA)	9.7
Gastrocnemius	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.5
Heart	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.3
Kidney	RNAlater	Polytron	AllPrep DNA/RNA Mini kit	9.4
Kidney	RNAlater	Polytron	Trizol (no DNA)	7.3
Kidney	LN2	Polytron	AllPrep DNA/RNA Mini kit	8.3
Kidney	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.4
Large intestine	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.7
Liver	LN2	Polyton	Trizol (no DNA)	6.2
Liver	LN2	Polytron	AllPrep DNA/RNA Mini kit	8.6
Liver	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.0
Lung	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	8.9
Mesenteric lymph nodes	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.6
Quadriceps	LN2	Polytron	Trizol (no DNA)	9.8
Quadriceps	LN2	Polytron	AllPrep DNA/RNA Mini kit	9.7
Reproductive tract	RNAlater	Bead Homogenizer	AllPrep DNA/RNA Mini kit	9.2
Skin (dorsal and femoral)	RNAlater	Polytron	Trizol (no DNA)	8.4
Skin (dorsal and femoral)	LN2	Polytron	Trizol (no DNA)	6.6
Soleus (2)	RNAlater	Polytron	AllPrep DNA/RNA Mini kit	9.8
Soleus (1)	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	8.7
Spleen	LN2	Polytron	AllPrep DNA/RNA Mini kit	6.8
Spleen	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	6.1
Thymus	LN2	Bead Homogenizer	AllPrep DNA/RNA Mini kit	7.7
Tibialis anterior	RNAlater	Polytron	Trizol (no DNA)	8.8

QC : iSeq and normalization of pools

Protocol that was selected for library prep involves manual rRNA depletion prior the automation steps.

Using iSeq to validate pool normalization and rRNA%



Library size distribution and dimer contamination checked on TapeStation

Multiplexing libraries by volume

1. Using clustering% to define normalization values for next pool
2. Validating all samples above rRNA criteria (0-10%)

Normalized pool multiplexing

1. Validating normalization
2. Using (Clustering% - rRNA%) to predict how many read each sample will have from NovaSeq

GeneLab Recommended Sequencing Standards

GeneLab suggests the following sequencing parameters to best capture important signals in RNA-sequencing data.

	Representative genome	Transcriptome complexity	Ribodepleted RNA-seq	Poly-A enriched RNA-seq
Human	*H. sapiens* GRCh38.p12	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Rodents (mouse and rat)	*M. musculus* GRCm38.p6	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Fruit fly	*D. melanogaster* Release 6 plus ISO1 MT	Medium	40 M clusters, 150 bp PE	30 M clusters, 150 bp PE
Worms	*C. elegans* WBcel235	Medium	40 M clusters, 150 bp PE	30 M clusters, 150 bp PE
Plants	*A. thaliana* TAIR10.1	High	60 M clusters, 150 bp PE	40 M clusters, 150 bp PE
Fungi	*S. cerevisiae* S288C R64	Low	20 M clusters, 150 bp PE	13 M clusters, 150 bp PE
Bacteria	*E. coli* str. K-12 substr. MG1655	Low	20 M clusters, 150 bp PE	13 M clusters, 150 bp PE

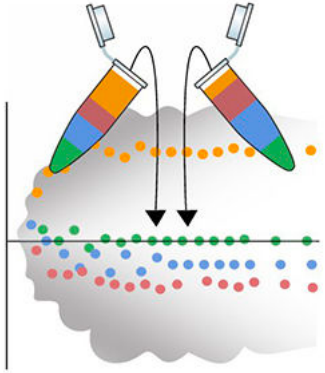
Spike-in Controls. At a minimum, GeneLab recommends that all sequencing samples include ERCC Spike-in Mix 1 in all samples. This will allow assessment of the dynamic range within each sample. In addition, in experiments focused upon comparing 2 levels of a given factor (spaceflight vs. ground control), GeneLab recommends including Mix 1 in the level A samples (spaceflight) and Mix 2 in the level B samples (ground control). This allows direct assessment of the available power in the data. In all cases, these Mixes should be added to samples before library preparation and sequencing. These Spike-in standards are available from Thermo Fisher: Mix 1 (https://www.thermofisher.com/order/catalog/product/4456740?SID=srch-srp-4456740), Mix 1 and 2 (https://www.thermofisher.com/order/catalog/product/4456739?SID=srch-srp-4456739).

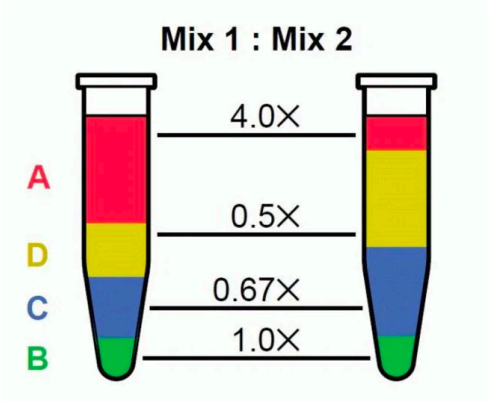
Library preparation and sequencing standard. GeneLab recommends that all sample pools to be sequenced include at least one replicate of the SEQC Universal Reference RNA available from Agilent (https://www.agilent.com/en/product/gene-expression-microarray-platform/gene-expression-microarray-kits-reagents/gene-expression-universal-reference-rnas-228491). This RNA should be included as a sample during each library preparation batch.

RNA-sequencing quality controls

GeneLab utilizes two types of sequencing controls:

1. Internal control spiked into each sample. GeneLab utilizes ERCC Mix 1 and Mix 2. Typically Mix 1 is spiked into “spaceflight” samples and Mix 2 is spiked into other groups to allow as many comparisons to the ”spaceflight” group as possible.





2. SEQC UHRR UMRR RNA standards:

UMRR - Universal Mouse Reference RNA is composed of total RNA from 11 mouse cell lines.
UHRR - Universal Human Reference RNA is composed of total RNA from 10 human cell lines.

Controls are includes with each RNA-seq library generation pool of 48 samples.

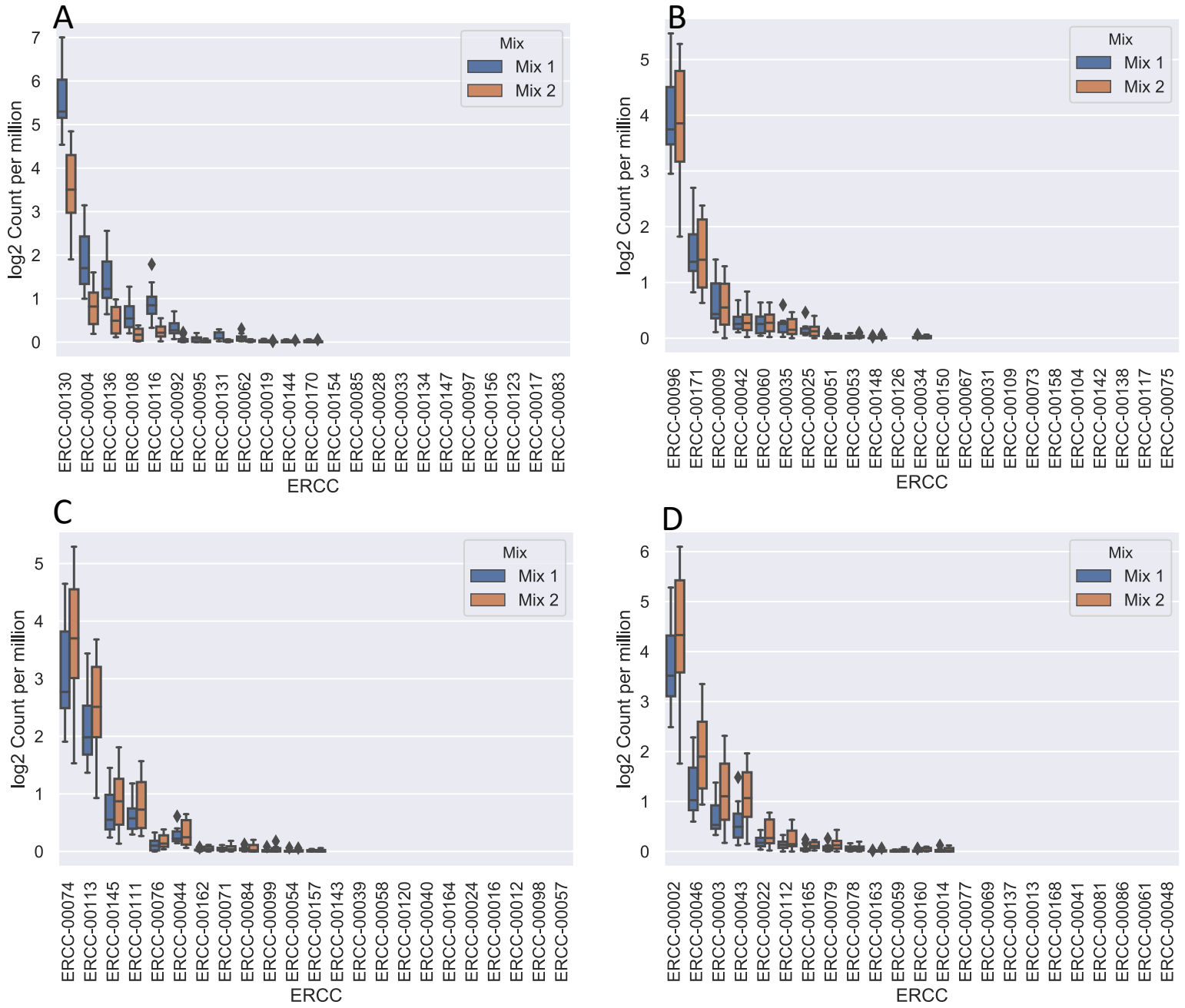
GLDS-239: Transcriptomic analysis of femoral skin from mice flown on the MHU-2 mission

Group A: Mix 1 present at 4x the concentration as Mix 2 (log2 FC of 2). 30,000 - 0.007 attomoles/ul.

Group B: Mix 1 present at 1x the concentration as Mix 2 (log2 FC of 0). 15,000 – 0.14 attomoles/ul.

Group C: Mix 1 present at 0.67x the concentration as Mix 2 (log2 FC of -0.58). 22,500 – 0.014 attomoles/ul.

Group D: Mix 1 present at 0.5x the concentration as Mix 2 (log2 FC of -1). 30,000 – 0.014 attomoles/ul.



Inclusion of UMI (Unique Molecular Identifier) in barcode

GeneLab utilizes xGen Dual Index UMI adapters from IDT
These UMI's allow exact quantitation of PCR duplicates
This is important, as RNA-seq data is relatively low-complexity and typically includes a fraction of apparent duplicates.

Figure 1. xGen Dual Index UMI Adapter.

Total	R1 R2 UMI dup	R1 R2 dup	R1 dup	R2 dup	%nR1 R2 UMI dup	% R1 R2 dup	% R1 dup	% R2 dup
1000000	2475	45815	77942	80063	0.2475	4.5815	7.7942	8.0063
5000000	62945	383977	630918	637282	1.2589	7.67954	12.61836	12.74564
50000000	5514952	10161055	15678069	14442464	11.029904	184.245575	154.295681	92.1188955
10000000	246295	988501	1605149	1591874	2.46295	9.88501	16.05149	15.91874
20000000	960116	2639874	4210820	4066633	4.80058	13.19937	21.0541	20.333165
1500000	5687	77768	131238	134495	0.379133333	5.18453333	8.7492	8.96633333
500000	640	18272	31975	33032	0.128	3.6544	6.395	6.6064
68000000	**9721538**	**16094465**	**24487015**	**22158598**	**14.29637941**	**23.6683309**	**36.0103162**	**32.5861735**

RNA-seq quality metrics

While these are under development metrics may include:

Pre-pipeline (from FASTQC):

- Sequence counts
- Sequence quality histograms
- Per sequence quality scores
- Per sequence GC content
- Per base N content
- Sequence length distribution
- Sequence duplication levels
- Overrepresented sequences
- Adapter content

Post-pipeline:

% rRNA

% uniquely mapped

% true duplicates (from Unique Molecular Identifier)