SINGLE-CELL DATA

Lee Panter

under advisement of

Audrey E Hendricks, PhD

University of Colorado Denver

Monday, November 11th, 2019

Department of Mathematical
& Statistical Sciences

UNIVERSITY OF COLORADO **DENVER**

- What is Single Cell Data?
- How is Single Cell Data Analyzed?
- Lupus Case Study of Single Cell Analysis
- More Information?

Data About Cells : Sample | Selection | Measurement(s)

- Sample Types: blood samples, mouth/cheek swabs, biopsies, hair follicle samples
- Selection Methods: (Re)combination (BULK), Isolation (Single Cell) (SC)
- Measurement(s): HBA1C, Flow Cytometry, Mass Spectrometry, RNA Sequencing

SELECTION METHOD	PRO	CON
BULK	• Accurate estimate of average expressions • Cleaner, more reproducible data • Cheaper	• Assumes average expression is informative of desired feature • Difficult to isolate cell-cell variation and identify cell-specific traits
SINGLE CELL	• Useful for assessing cell-cell variations and identifying rate traits • Distributional properties other than average (skewness, kurtosis,...etc) associable with desired features	• Difficult and expensive • Individual cell information can be noisy • Messy data, prone to measurement biases, and batch effects

[1] [2]

Seurat – R Package – Guided Clustering Tutotorial [3]

Resources for:

- Data Management
  - Quality Control, Identification of Highly Variable Features
- Visualization
  - Linear Dimensional Reduction - Principle Components Analysis (PCA)
  - Non-linear Dimensional Reduction:
    - Uniform Manifold Approximation and Projection (UMAP)
    - T-distributed Stochastic Neighbor Embedding (t-SNE)
- Unsupervised Learning
  - Cellular Clustering (PCA Space, Differentially Expressed Features)

Details on Quality Control included in this presentation, for other features visit: `https://satijalab.org/seurat/`

Seurat Tutorial Data [4]

- Single-Cell RNA Sequencing (scRNA-SEQ) expression profiles
- Source: 10X genomics
- 2,700 Single-Cell observations on 32,738 Features (i.e. genes)
- $\sim$ 88 million values

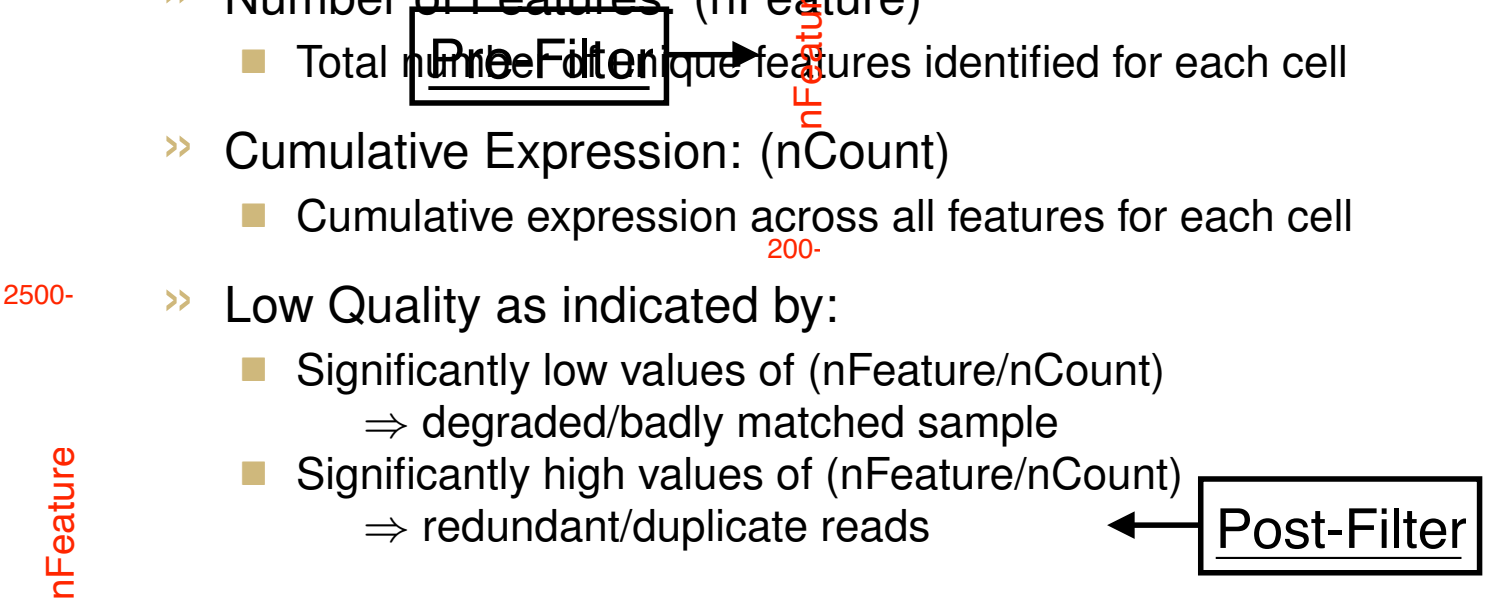
RNA Sequencing (on each cell)

- Expression Profiles: which genes are being expressed at any given moment?
- Cell type, state, environment [5]
- Count values correspond to: "number of reads overlapping a given feature" [6]

Seurat Tutorial nFeature Threshold

Quality Control Metrics: nCount & nFeature

- Number of Features and Cumulative Expression Count
  - Number of Features: (nFeature)
    - Total number of unique features identified for each cell
  - Cumulative Expression: (nCount)
    - Cumulative expression across all features for each cell
  - Low Quality as indicated by:
    - Significantly low values of (nFeature/nCount)
      ⇒ degraded/badly matched sample
    - Significantly high values of (nFeature/nCount)
      ⇒ redundant/duplicate reads

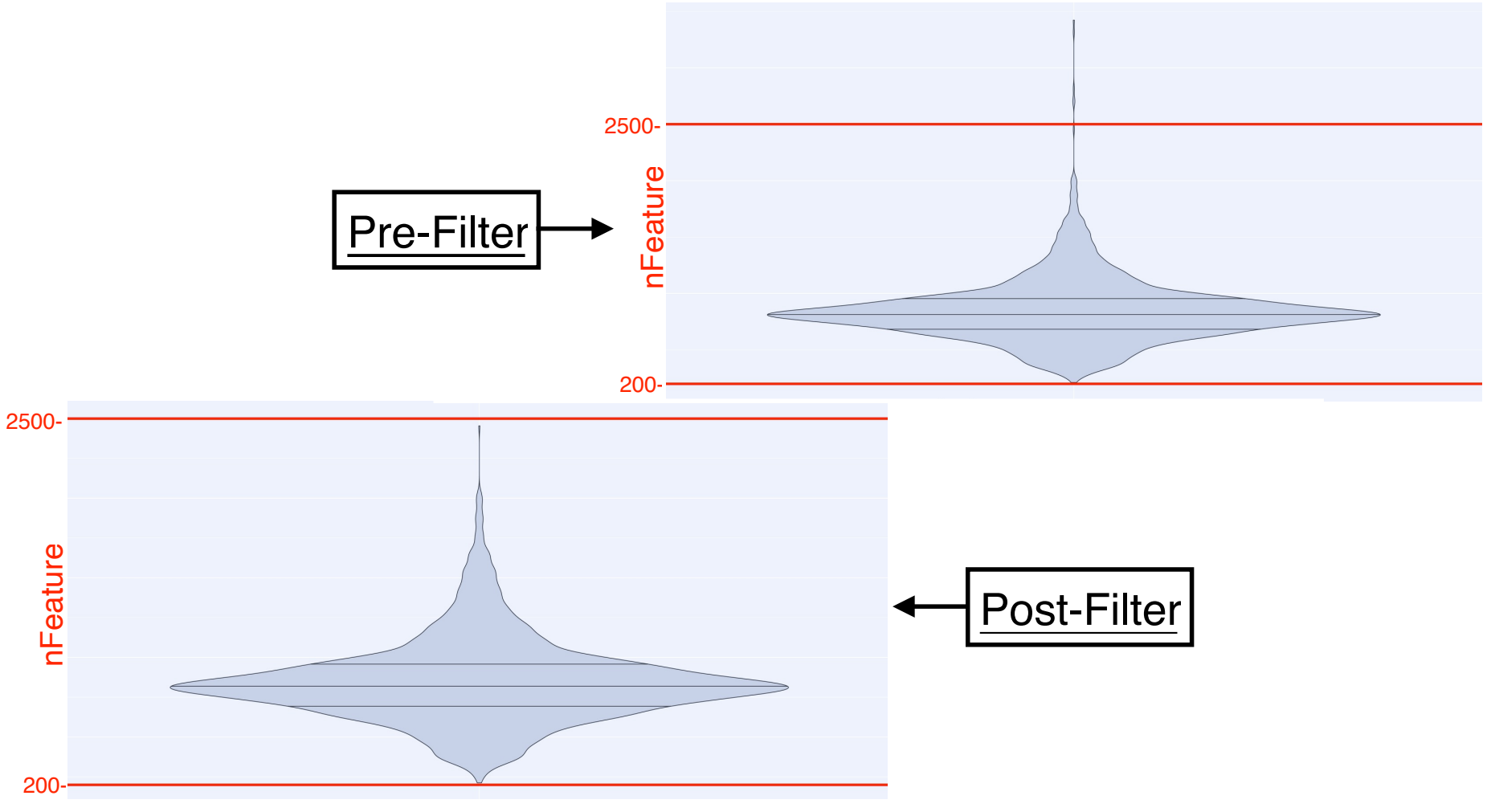


Seurat Tutorial Imposed QC Thresholds

200-

- 200 < nFeature < 2,500

Seurat Tutorial nFeature Threshold



Quality Control Metric: Perct.Mt

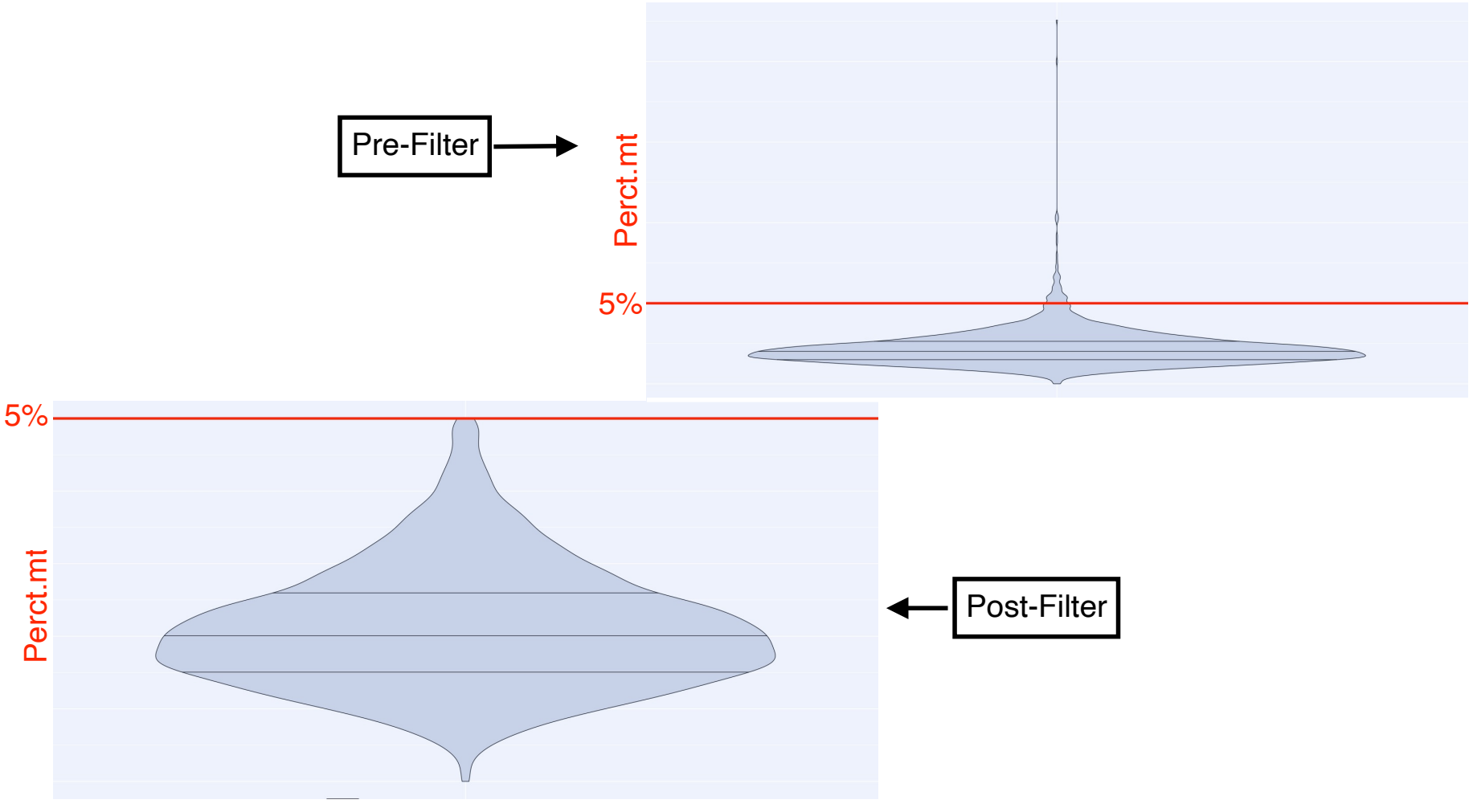
- Percentage Mitochondrial: Perct.Mt
  - Percentage of cumulative expression that maps to the mitochondrial genome
  - $\sim$ 37 features, all feature designations marked: "MT-"
  - Normalized with respect to the cumulative cellular expression (nCount)

- Low Quality as indicated by:
  - high expression of mitochondrial features
    $\Rightarrow$ dying/dead cells, extensive mitochondrial contamination.

Seurat Tutorial Imposed QC Thresholds

- Perct.Mt < 5%

Seurat Tutorial Perct.Mt Threshold



Case Study in Single Cell Data: Lupus Study [7]

Data About Cells : Sample | Selection | Measurement(s)

The Sample:

- 27 human subjects from 10 clinical sites across the U.S
- Kidney Biopsies and Urine Samples
- Samples were frozen, shipped, & processed centrally

Selection Method:

- 11-color Flow Cytometry: 23 variables per cell
- 9,560 total single-cell observations from the 27 samples

Measurement(s):

- scRNA-SEQ expression profiles for all 9,560 observations
- 38,354 unique expression markers each observation
- Data available: `https://www.immport.org/shared/study/SDY997`

Initial Summary of Lupus Data

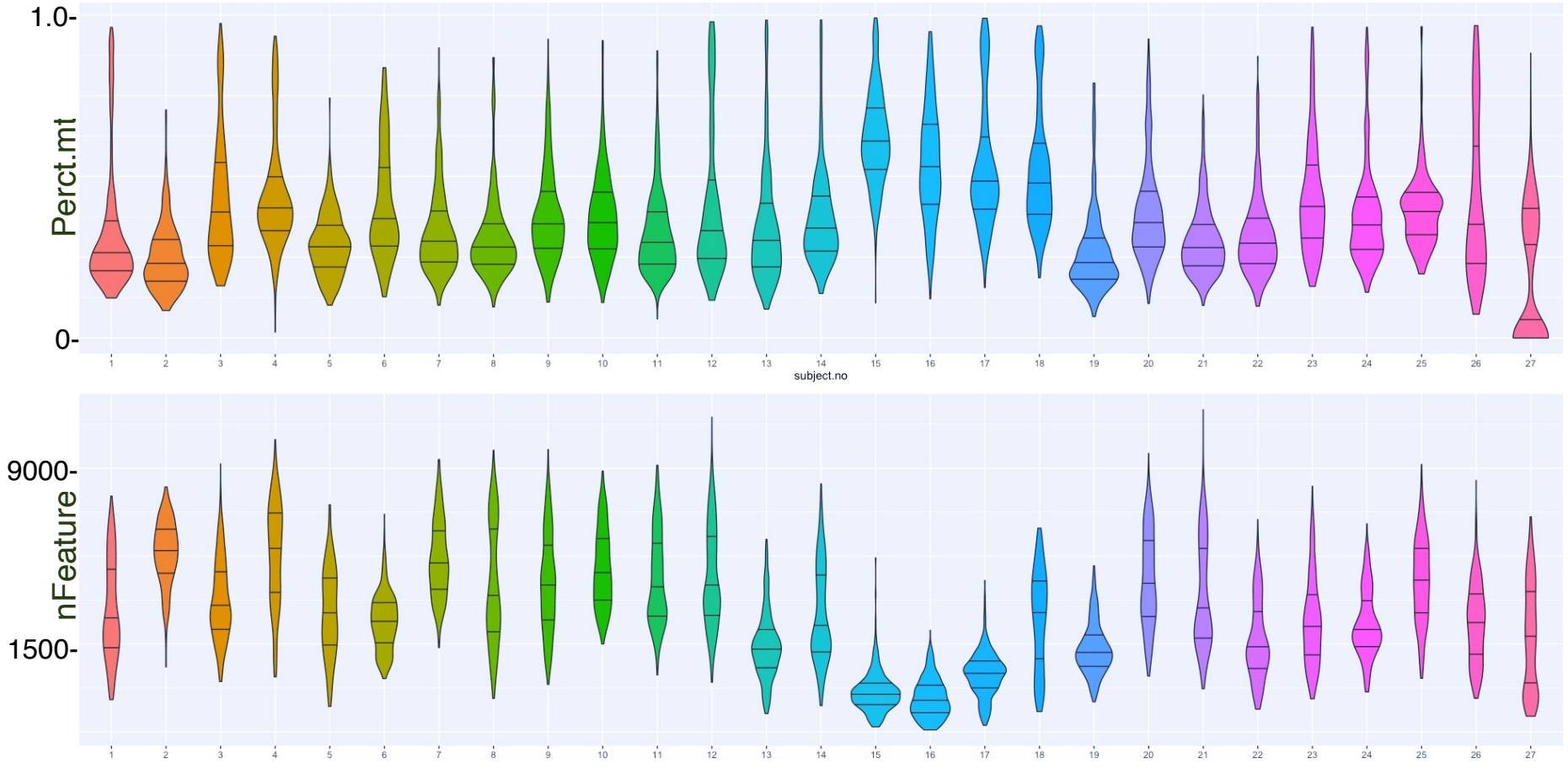


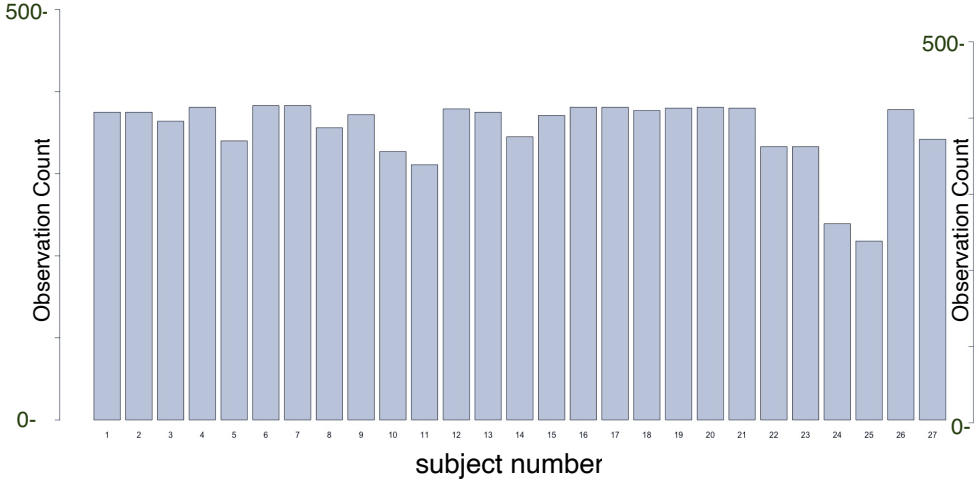
Figure: Perct.mt and nFeature arranged by subject

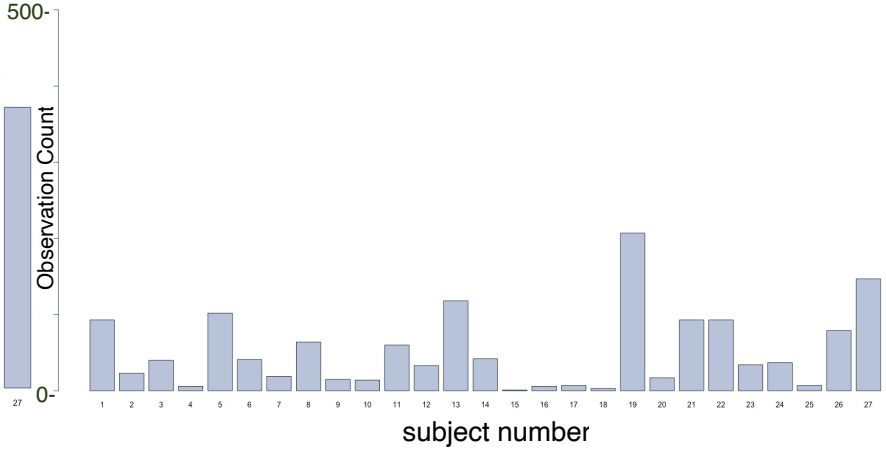
Analysis of Lupus Paper QC Thresholds

QC Measure	Seurat Tutorial Value	Lupus Paper Values
nFeature	$200 < \text{nFeature} < 2{,}500$	$1{,}000 < \text{nFeature} < 5{,}000$
Perct.mt	$\text{Perct.mt} < 5\%$	$\text{Perct.mt} < 25\ \%$

Subject Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Observations Before Filter	375	375	364	381	340	383	383	356	372	327	311	379	375	345	371	381	381	377	380	381	380	333	333	239	218	378	342
Observations After Filter	93	23	40	6	102	41	19	64	15	14	60	33	118	42	1	6	7	3	207	17	93	93	34	37	7	79	147

Average Before Filter	354
Average After Filter	52





Lupus Paper QC Threshold for Perct.Mt

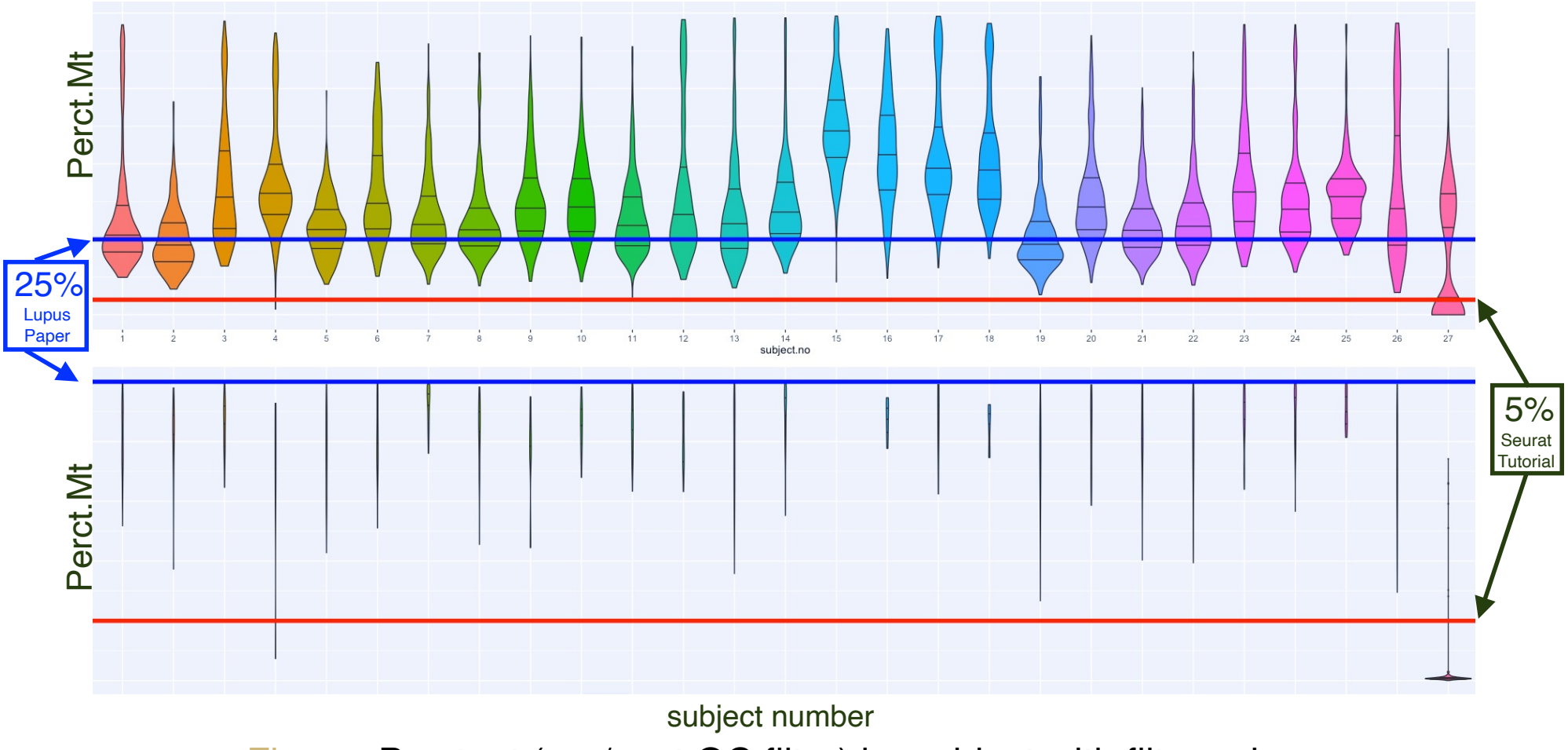


Figure: Perct.mt (pre/post QC filter) by subject with filter values

Lupus Paper QC Threshold for nFeature

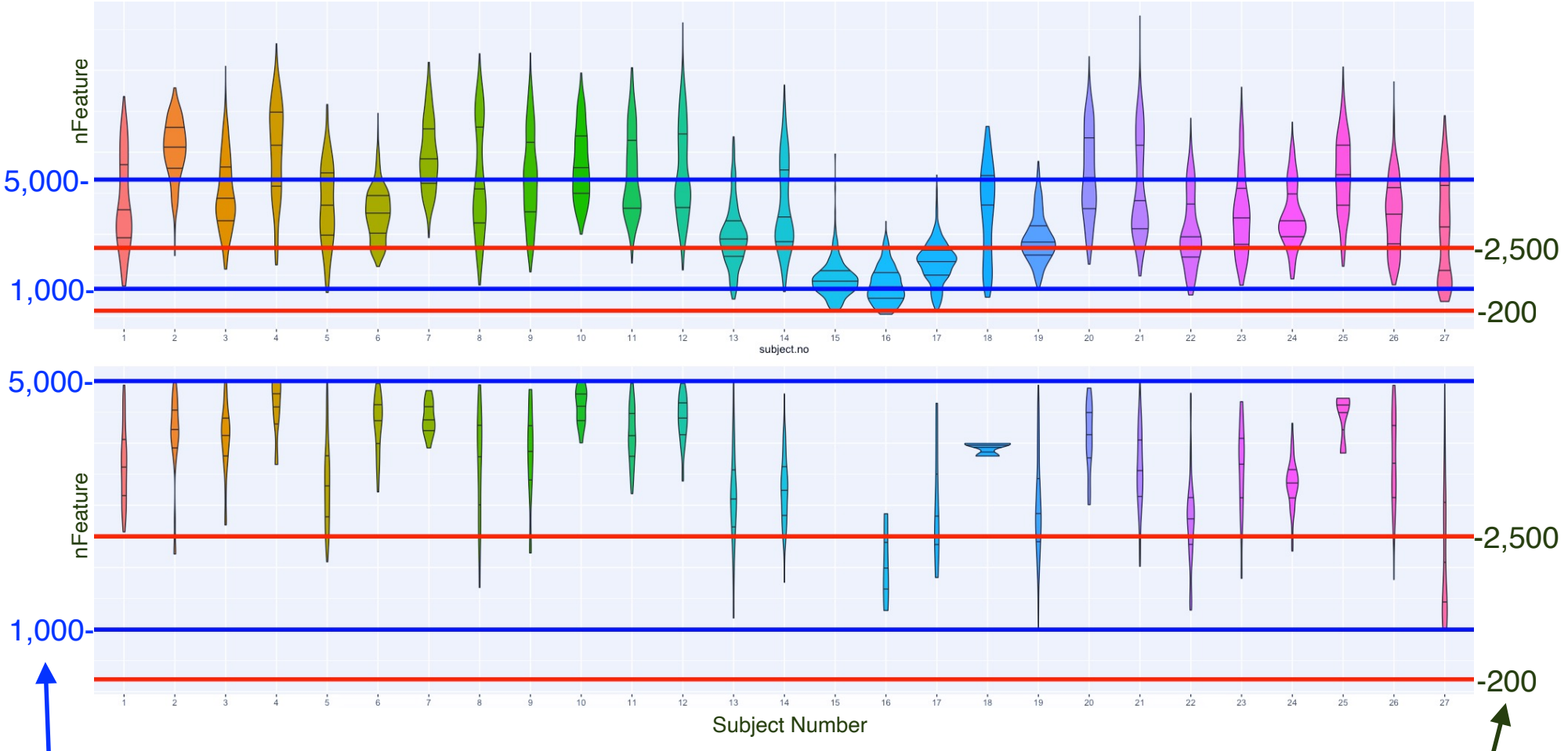


Figure: nFeature (pre/post QC filter) by subject with filter values

Lupus Paper

1,000 < nFeature < 5,000

Seurat Tutorial

200 < nFeature < 2,500

Conclusion: Filters imposed by the Lupus paper cause information imbalance between sample sources

Possible Problems:

- Remaining observations are not reflective of sample
- Information weighted according to subject-level properties

What's the Solution?

- Get recommendations from an expert:
  - Relax QC threshold values still further
  - Restrict new data to B-cells
    - similar cells –> similar measurement values –> easier to identify reasonable QC thresholds

New QC Thresholds and Additional Filters

- 1,000 < nFeature < 5,000 (unchanged)
- Perct.mt < 60%
- New Filter: B-Cells Only

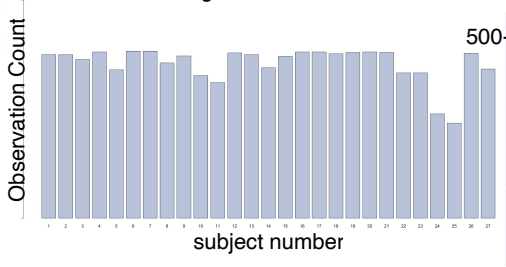
Analysis of Recommended QC Thresholds

QC Measure	Seurat Tutorial Value	Lupus Paper Values	New Recommendations
nFeature	$200 < \text{nFeature} < 2{,}500$	$1{,}000 < \text{nFeature} < 5{,}000$	$1{,}000 < \text{nFeature} < 5{,}000$
Perct.mt	$\text{Perct.mt} < 5\%$	$\text{Perct.mt} < 25\%$	$\text{Perct.mt} < 60\%$

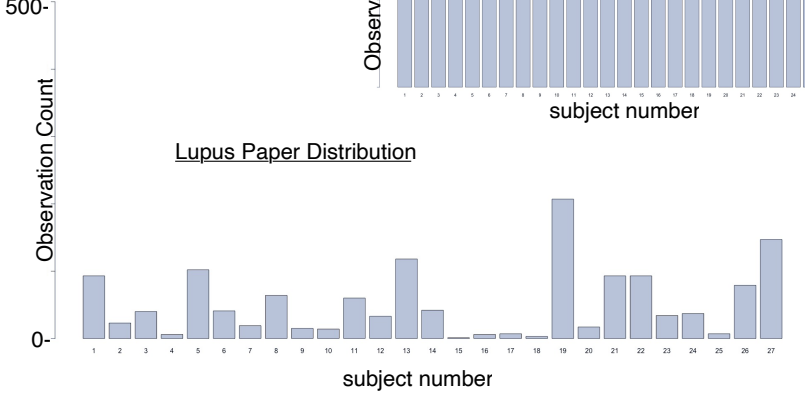
Subject Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Observations Before Filter	375	375	364	381	340	383	383	356	372	327	311	379	375	345	371	381	381	377	380	381	380	333	333	239	218	378	342
Observations After Filter	93	23	40	6	102	41	19	64	15	14	60	33	118	42	1	6	7	3	207	17	93	93	34	37	7	79	147
Observations After New Filter	0	0	0	0	58	86	32	0	31	21	107	107	0	100	25	0	122	0	127	75	0	87	0	79	0	53	0

Average Before Filter	354
Average After Filter	52
Average After New Filter	41

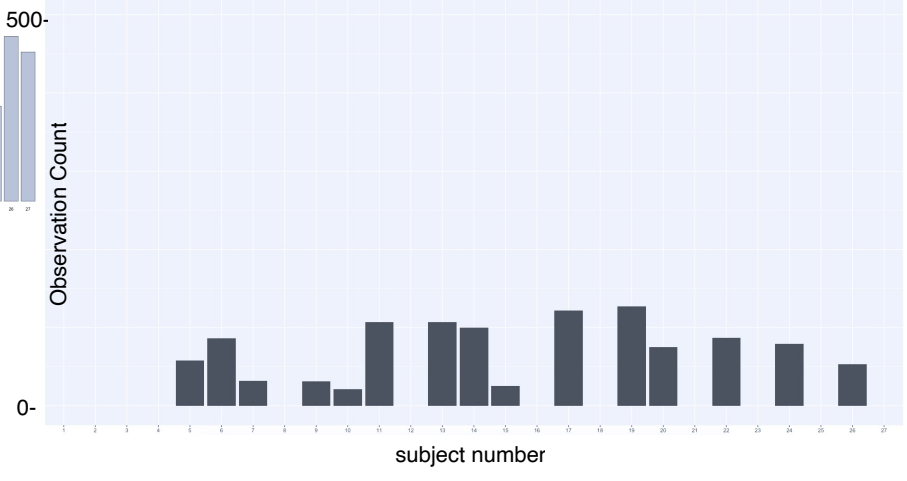
Original Distribution



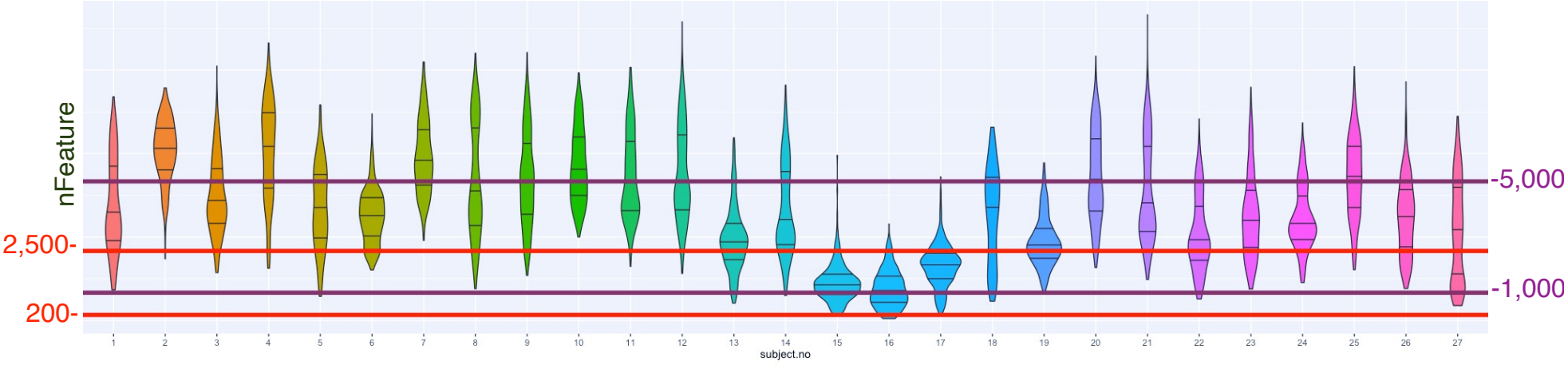
Lupus Paper Distribution

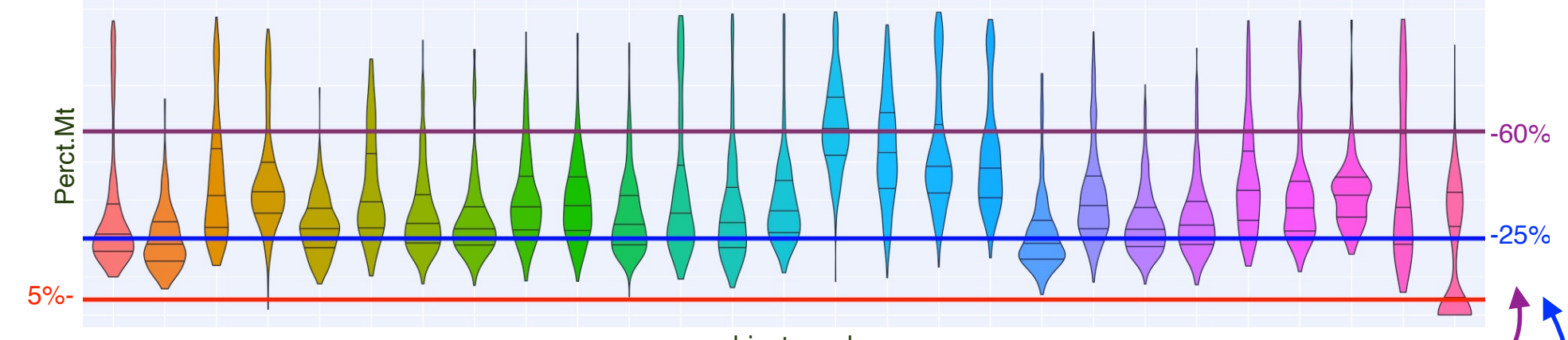


Recommended Value Distribution



Recommended Threshold Comparisons





subject number

Seurat Tutorial

200 < nFeature < 2,500

Perct.Mt < 5%

Lupus Paper

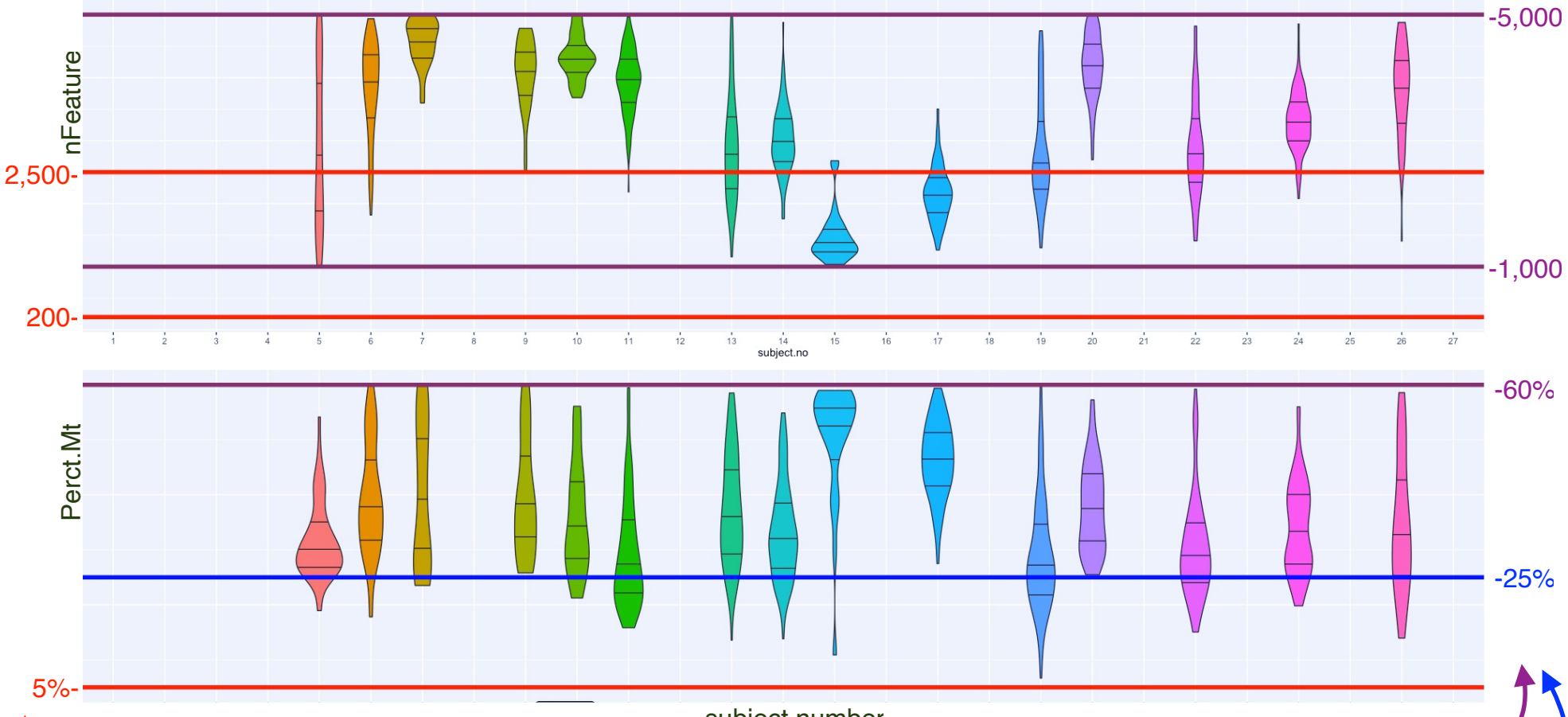
1,000 < nFeature < 5,000

Perct.Mt < 60%

Adopted Values

Perct.Mt < 25%

Finalized Data with Threshold Values



Seurat Tutorial

200 < nFeature < 2,500

Perct.Mt < 5%

Lupus Paper

1,000 < nFeature < 5,000

Perct.Mt < 60%

Adopted Values

Perct.Mt < 25%

Lessons Learned

- Single-Cell data is messy-Seurat package as a guide-may need to alter QC threshold parameters
- Having an expert to consult is necessary
- If a threshold change is impossible to avoid, make changes to increase statistical power elsewhere.

References I


James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim.
The promise of single-cell sequencing.
Nature methods, 11(1):25, 2014.


Yong Wang and Nicholas E Navin.
Advances and applications of single-cell sequencing technologies.
Molecular cell, 58(4):598–609, 2015.


Satija lab.
`https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html`.
(Accessed on 11/07/2019).


10x genomics: Resolving biology to advance human health.
`https://www.10xgenomics.com/`.
(Accessed on 11/07/2019).

References II

Gene expression profiling - wikipedia.
https://en.wikipedia.org/wiki/Gene_expression_profiling.
(Accessed on 11/09/2019).

Counts vs. fpkms in rna-seq.
https://www.cureffi.org/2013/09/12/counts-vs-fpkms-in-rna-seq/.

(Accessed on 11/09/2019).

Arnon Arazi, Deepak A Rao, Celine C Berthier, Anne Davidson, Yanyan Liu, Paul J Hoover, Adam Chicoine, Thomas M Eisenhaure, A Helena Jonsson, Shuqiang Li, et al.
The immune cell landscape in kidneys of lupus nephritis patients.
bioRxiv, page 363051, 2018.

Interested In Learning More?

- My project GitHub: `https://github.com/leepanter/RBC`
- Email: `lee.panter@ucdenver.edu`
- Presentation Links
  - Seurat Tutorial: `https://satijalab.org/seurat/`
  - Lupus Data: `https://www.immport.org/shared/study/SDY997`