A study of the cellular and environmental context of Barrett's esophagus



David T. Severson

St. Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2017

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Benjamin Schuster Böckler and Professor Xin Lu for their extensive scientific and career advice over the last four and a half years. Without them this work would not exist, and I would not be the scientist I am today. Xin helped guide me toward a completely computational career as I struggled with wet lab experimentation early in my DPhil and it has proven to be a wise path. I have adored my involvement in applying single-cell analysis to the Barrett's problem, and Xin provided the opportunity to be a part of such an exciting project. And I could not have been successful without the extensive day to day mentorship of Benjamin. From the mundane task of learning programming to the more nuanced and challenging work of wrestling with the design of a new method, Benjamin has been an essential mentor and resource during the DPhil process. I also want to express extensive gratitude to both Benjamin and Xin for their patience and understanding concerning my unconventional completion of the D. Phil on account of my mother's illness, my wife's employment at Brown University, and my application to medical school this year.

I want thank my colleagues and friends, Richard Owen and Michael White. Not only were they crucial to the materialization of this work, but Richard and Mike helped make even the hardest moments of the degree fun. In the dingy basement of the Old Road Research Campus Building, lengthy conversations with Marketa Tomkova about science and life were important for both my scientific development and my sanity. I would also like to thank Felix Zhou for all of his effort and discussion involving the work surrounding EBV and nasopharyngeal carcinoma, and for his helpful input as I navigated the development of BEARscc with Benjamin. I owe Andy Roth and Mary Muers a hearty thank you for their crucial input on the manuscript introducing BEARscc. In general, I would like to thank all the members of the Xin Lu lab for their support and companionship throughout the D.Phil. I also owe a great deal of gratitude to all of the collaborators with which I have worked during my time at the University of Oxford; needless to say, none of my work would be complete their efforts.

The Teddy Hall MCR was a crucial pillar in my social well-being and happiness, and provided some of the best experiences of my life. I want to thank members of the community who supported me during my times in leadership roles, and for

creating such a thrilling culture and community in which I was fortunate enough to live for four years. In particular, I want to thank Harry Clifford, Graham Baird, Lisa Pochmalicki and Antonin DeLaever for profound friendship that sustained me during the degree. I also want to thank the Teddy Hall Boat Club for providing a social and physical avenue to relax and decompress from work.

I want to thank the Bueno lab and Dr. Raphael Bueno for their patience and understanding as I have worked to complete this degree and related manuscripts while employed at Brigham and Women's hospital.

Most importantly, I want to thank my wife, Laura Hawkins for her companionship and support. I could not have completed the degree without her, and I must admit that our marriage is the most precious achievement from my time at Teddy Hall and the University of Oxford. In similar fashion I would like to thank my three brothers and parents for unending love and support.

My mother's continued unconditional love and support in the face of illness has amazed me. I would like to dedicate this thesis to her as a thank you for raising me in a loving and nurturing environment that fostered the success with which I have been blessed thus far and with which I hope to be blessed in the future.

Abstract

Barrett's esophagus is a pre-malignant lesion leading to esophageal adenocarcinoma, where presence of Barrett's carries a seven fold increase in risk of developing cancer. The histology of Barrett's esophagus is diverse, containing squamous, gastric, and intestinal features, and the origin of the lesion is still under considerable debate. Professor Lu's research group aims to define the molecular features underlying the Barrett's lesion and elucidate the functional origin of the lesion. Because pathogens and commensals have previously been reported to be important in other cancers of the upper gastro-intestinal tract, we also aim to explore the role of microbes and viruses in Barrett's esophagus. We sequenced 3069 single cell RNA samples from endoscopic biopsies of the Barrett's lesion and relevant normal tissue in the upper gastro-intestinal tract and positive and negative controls. In order to draw meaningful conclusions from these single cell RNA data concerning the functional origin of Barrett's, we needed to identify robust relationships between single cell observations across tissue types. However, technical variance is a major confounding factor in single-cell RNA sequencing and could distort any such observed relationships. Therefore, I developed BEARscc, a tool that evaluates cluster robustness to noise by simulating experiment-specific technical replicates. I demonstrate that the tool improves the unsupervised classification of cells and aids the interpretation of single-cell RNA-seq experiments. I apply BEARscc to our single cell RNA-sequencing cohort and find striking similarity between the esophagus submucosal glandular ducts and a subset of Barrett's cells with undifferentiated characteristics. Using BEARscc, I also provide a robust description of the substructure of normal esophageal, gastric, and duodenal tissues. In order to characterize any microbial or viral presence in the microenvironment of Barrett's, I developed a pipeline, poshTitan, to identify pathogens and commensals in next generation sequencing data. I demonstrate that the poshTitan identifies pathogens with known associations in nasopharyngeal squamous cell carcinoma whole tissue RNA-sequencing and gastric adenocarcinoma whole genome sequencing data. I apply the poshTitan to whole tissue and single cell RNA-sequencing samples of Barrett's esophagus and normal controls and find no clear pattern of pathogen or commensal association with disease.

Contents

List of Figures **x**

List of Tables **xv**

1 Introduction **1**
1.1 The problem in context 1
1.2 Motivation 11
1.3 Contribution and Scope 13

2 BEARscc models noise to reveal cluster robustness **17**
2.1 Introduction 18
2.2 Data and methods 21
2.2.1 Data 21
2.2.2 Clustering of counts data 22
2.2.3 Estimation of cluster number k 22
2.2.4 Gene sampling for benchmarking 23
2.3 Algorithm description and implementation 23
2.3.1 BEARscc workflow 23
2.3.2 Algorithmic generation of simulated technical replicates 26
2.3.3 Modelling noise from spike-ins 27
2.3.4 Inference of transcript dropout distributions using spike-ins . . 29
2.3.5 Computation of consensus matrix 31
2.3.6 Computation of BEARscc cluster metrics 32
2.3.7 Computation of BEARscc cell metrics 32
2.3.8 Implementation 32
2.4 Optimizing BEARscc to accurately simulate technical replicates . . . 33
2.4.1 Pseudocode illustrating the variations in implementation for the BEARscc algorithm 35
2.4.2 Gene-wise implementations of BEARscc are superior to global implementations 36
2.4.3 A single drop-out function most accurately models observed drop-out number 43

2.5 BEARscc accurately models technical noise in single cell RNA-seq data 46
2.6 The importance of normalization in clustering analysis 48
2.7 Performance of BEARscc combined with different clustering algrotihms 55
2.8 Choosing the number of simulated replicates 60
2.9 BEARscc identifies cells and clusters that are robust to noise 62
2.9.1 An illustration in murine intestine single cell samples 62
2.9.2 An illustration in murine brain single cell samples 67
2.10 Benchmarking BEARscc against simple gene sampling 70
2.11 Conclusions and discussion 72

3 Single cell RNA-sequencing reveals insights in Barrett's esophagus 75
3.1 Introduction 76
3.2 Experimental design 79
3.2.1 Endoscopic biopsies of four upper GI tract tissues are collected from six patients 80
3.2.2 Endoscopic biopsies are FACS sorted into single cells 81
3.2.3 Controls for batch and successful sequencing of transcriptomes 81
3.2.4 Description of samples sequenced with single cell RNA-sequencing 81
3.3 Data and methods 84
3.3.1 Patient enrollment and endoscopic biopsies sampling 84
3.3.2 Isolation of single cells from endoscopic biopsies 85
3.3.3 Single cell RNA-sequencing 86
3.3.4 Bulk RNA-sequencing 86
3.3.5 Library normalization 86
3.3.6 Differential expression analysis to identify marker genes 87
3.3.7 Immunohistochemistry and immunofluorescence staining 87
3.4 Duplicate removal does not significantly alter biological signal 88
3.4.1 Duplication is high in our dataset 89
3.4.2 Duplication rates do not correlate with library complexity . . . 90
3.4.3 Duplicate removal has a small impact on most genes 93
3.4.4 Gene-wise duplication rates are independent of gene length and disproportionately high in highly expressed genes 94
3.4.5 Duplicate removal does not affect biological signal 96
3.4.6 Duplicates were removed before further analysis of our data . . 101
3.5 5' or 3' bias was not present in our data. 101
3.6 Mapping metrics and QC of single cell RNA-sequencing libraries . . . 102
3.6.1 Mapping metrics are not effective indicators of library quality in our data 102
3.6.2 ERCC spike-in concentration in our samples 106
3.7 Mitochondrial and rRNA expression in our samples 111

3.7.1 Nuclear rRNA gene expression is low in our samples 111
3.7.2 Mitochondrial gene expression is high in our libraries 114
3.8 High quality libraries were obtained from endoscopic biopsies. 117
3.8.1 Transcriptome fragment count determines high quality libraries 117
3.8.2 Transcriptome fragment count and library complexity 120
3.8.3 ERCC content is higher in low transcriptome fragment libraries 125
3.8.4 Feature selection for clustering analysis 126
3.8.5 1821 samples have high quality transcriptomes 127
3.9 Overview of BEARscc analysis of Barrett's experiment libraries . . . 129
3.10 Determining clusters for biological interpretation 131
3.10.1 Initial SC3 cluster number parameters in individual tissues . . 131
3.10.2 Initial SC3 cluster number parameter across tissue types 136
3.10.3 Final clusters are determined by BEARscc 137
3.11 BEARscc identifies substructure in individual gastro-intestinal epithelia 143
3.11.1 Summary of subcellular structure in upper GI epithelia 144
3.11.2 Understanding the observed cellular structure 145
3.12 Interpreting the relationships between cells of different tissue types. . 147
3.12.1 The biology of the 12 BEARscc curated clusters 148
3.12.2 Barrett's esophagus is functionally similar to esophageal submucosal glandular duct cells 154
3.12.3 OLFM4 and LEFTY1 are expressed in Barrett's esophagus . . 158
3.12.4 ITLN1 and SPINK4 genes mark early goblet cells 160
3.13 Conclusions and discussion 162

4 poshTitan: Detecting pathogens in next-generation sequencing data 167
4.1 Introduction 168
4.2 Data and methods 170
4.2.1 Gastric adenocarcinoma whole genome sequencing data 170
4.2.2 Nasopharyngeal carcinoma RNA sequencing data 170
4.3 poshTitan workflow and implementation 170
4.3.1 Pipeline description 170
4.3.2 Implementation 174
4.4 poshTitan accurately detects pathogens in simulated data 175
4.4.1 Performance of the Kraken fast alignment step 176
4.4.2 poshTitan is effective in simulated libraries 178
4.5 poshTitan detects *EBV* and *H. Pylori* in gastric cancer 180
4.5.1 Landscape of pathogens in gastric cancer 181
4.5.2 poshTitan reliably detects EBV and identifies *H. Pylori* in gastric cancer 187

4.5.3 Human herpesvirus 5 (*cytomegalovirus*) was identified in two patients' tumor samples . . . 188
4.6 poshTitan identifies EBV as ubiquitous in nasopharyngeal carcinoma 192
4.7 Pathogens do not appear to be important in Barrett's esophagus . . 196
4.7.1 No pathogen association with Barrett's esophagus is observed in whole tissue libraries . . . 196
4.7.2 No pathogen association with Barrett's is observed in single cell libraries . . . 200
4.8 Conclusions and discussion . . . 204

5 Conclusions and Discussion 205

6 Afterword 211
6.1 Beyond the scope of this work . . . 211
6.2 Relevant publications . . . 215

Appendices

A Supplementary Figures, Tables, and Software 221
A.1 Supplementary Figures . . . 222
A.1.1 BEARscc benchmarking . . . 222
A.1.2 Single cell cohort description . . . 225
A.1.3 Esophageal cluster interpretation . . . 230
A.1.4 Gastric cluster interpretation . . . 238
A.1.5 Duodenal cluster interpretation . . . 246
A.1.6 Barrett's tissue cluster interpretation . . . 256
A.1.7 All QC-filtered cell cluster interpretation . . . 266
A.1.8 Further analysis of all QC-filtered cells . . . 281
A.1.9 Single cell pathogen detection . . . 285
A.2 Supplementary Tables . . . 287
A.2.1 Single cell Barrett's cohort descriptions . . . 287
A.2.2 BEARscc metrics and differential expression exclusion . . . 288
A.2.3 Differentially expressed genes across the esophageal clusters . . 291
A.2.4 Differentially expressed genes across the gastric clusters . . . 315
A.2.5 Differentially expressed genes across the duodenal clusters . . . 320
A.2.6 Differentially expressed genes across Barrett's tissue clusters . 328
A.2.7 Differentially expressed genes in clusters from all QC-filtered cells 334
A.2.8 Differential expression summary . . . 353
A.2.9 BEARscc metrics for clusters from all QC-filtered cells . . . 353
A.3 List of Software . . . 354

References 357

List of Figures

1.1.1 Structure of the normal esophagus. 4
1.1.2 Histological features of esophageal submucosal glandular ducts. . 4
1.1.3 Structure of the normal stomach gland. 6
1.1.4 Structure of the normal duodenum. 7

2.3.1 BEARscc algorithm workflow. 26
2.4.1 Coefficient of variance vs mean across implementations. 38
2.4.2 Logarithmic simulated vs observed noise coefficient of variance across implementations 39
2.4.3 Logarithmic simulated vs observed noise mean expression across implementations 41
2.4.4 Logarithmic simulated vs observed noise variance across implementations 42
2.4.5 Observed vs simulated noise dropout rates 44
2.4.6 Histogram of genes' dropout rates across implementations 45
2.5.1 BEARscc accurately models technical noise 47
2.6.1 BEARscc and normalization in RaceID2 52
2.6.2 BEARscc and normalization in BackSPIN 53
2.6.3 BEARscc and normalization in SC3 54
2.7.1 BEARscc and the batch experiment 56
2.7.2 BEARscc in brain and blanks experiment - RaceID2 and BackSPIN 57
2.7.3 BEARscc in brain and blanks experiment - SC3 58
2.7.4 Boxplot of brain and blanks library size 59
2.7.5 Filtered RaceID2-Bearscc clustering result 60
2.8.1 Determining the number of simulated technical replicates 61
2.9.1 BEARscc applied to murine intestinal data 63
2.9.2 R1 and R2 comparison in murine intestinal data 64
2.9.3 Robust clusters and cells in BEARscc applied to murine intestinal data 66
2.9.4 Cluster number plot for BEARscc applied to murine brain data . 68
2.9.5 BEARscc summary results for murine brain data 69
2.10.1 Benchmarking with adjusted rand index 71

3.2.1 Schematic of Barrett's scRNA-seq experimental design 82
3.4.1 Histogram of duplication rate across single cell libraries 90
3.4.2 Scatterplot of expressed gene number vs duplication rate by tissue 91
3.4.3 Scatterplot of expressed gene number vs duplication rate by batch 92
3.4.4 Histogram of Δ gene expression with duplicate removal 93
3.4.5 Histogram of Δ reported expression by length 95
3.4.6 Histogram of Δ reported expression by duplicates retained expression 96
3.4.7 PCA of preliminary single cell data - duplicates retained 99
3.4.8 PCA of preliminary single cell data - duplicates removed 100
3.5.1 Gene body bias heatmap 102
3.6.1 Histogram: Fraction of aligned fragments 103
3.6.2 Histogram: Number of genomic alignments 104
3.6.4 Scatterplot illustrating mapping rate vs sequencing depth 106
3.6.5 Boxplot of ERCC expression with drop-outs 107
3.6.6 Barplots of fraction dropouts per expected ERCC 108
3.6.7 Scatterplot of fraction ERCC spike-ins to sequencing depth 109
3.6.8 Distribution of ERCC content fraction across sample sources . . . 110
3.7.1 Distribution of rRNA content fraction across sample sources . . . 112
3.7.2 Barplot displays rRNA content fraction across whole tissue samples 113
3.7.3 Distribution of mitochondrial expression fraction across sample sources 115
3.7.4 Scatterplot displays mitochondrial vs ERCC content in blank samples 116
3.7.5 Barplot displays mitochondrial expression percentage across whole tissue samples 117
3.8.1 Fragments mapped to the transcriptome by sample type 118
3.8.2 Histogram of library size in fragments mapped to the transcriptome 119
3.8.3 Illustration of library complexity averaged for each sample type . 121
3.8.4 Illustration of library complexity 122
3.8.5 Minimum number of genes to reach 75% of expression 123
3.8.6 Expressed genes by threshold 124
3.8.7 ERCC spike-in observed concentration for included and excluded cells 126
3.8.8 Coefficients of variance vs mean expression 127
3.8.9 Barplot of quality control inclusion/exclusion by sample origin . . 129
3.10.1 SI distribution within individual tissues 133
3.10.2 BEARscc score at various k - duodenum 134
3.10.3 BEARscc score at various k - Barrett's 135
3.10.4 SI distribution across all tissue 136
3.10.5 BEARscc score at chosen k for individual tissues 138

3.10.6 BEARscc stability at chosen k for individual tissues 139
3.10.7 BEARscc promiscuity at chosen k for individual tissues 140
3.10.8 BEARscc metrics at chosen k in all-sample analysis 142
3.12.1 Barplot of tissue distribution across all-sample clusters - robust cells 149
3.12.2 t-sne plots of marker genes 153
3.12.3 Sankey diagram displays individual tissue and all-sample BEARscc clusters 156
3.12.4 Sankey diagram displays relationship between Barrett's and esophageal submucosal glandular cells 157
3.12.5 t-sne plots of stemness marker genes 160
3.12.6 t-sne plot of Goblet cell marker genes 162

4.3.1 poshTitan workflow 174
4.4.1 Barplot of simulated library composition 176
4.4.2 Scatterplot of false negative rate in simulated libraries by Kraken 177
4.4.3 Scatterplot of false positive rate in simulated libraries by Kraken . 178
4.4.4 Barplot of pathogens and commensals detected in simulated libraries 179
4.4.5 Boxplots of coverage and mapped reads in simulated libraries . . 180
4.5.1 Pathogens detected in gastric cancer samples - tumor and normal 183
4.5.2 Pathogens detected in gastric cancer samples - tumor or normal . 184
4.5.3 Boxplot of detected pathogen coverage in gastric patient libraries 185
4.5.4 Boxplot of detected pathogen mapping rate in gastric patient libraries 186
4.5.5 Barplot of human herpesvirus mapped reads in gastric samples . . 190
4.5.6 Barplot of human herpesvirus coverage in gastric samples 191
4.6.1 Pathogens detected in NPC samples 193
4.6.2 Boxplot of detected pathogen coverage in NPC patient libraries . 194
4.6.3 Boxplot of detected pathogen mapping rate in NPC patient libraries 195
4.7.1 Pathogens detected in whole tissue libraries from patients with Barrett's 197
4.7.2 Boxplot of detected pathogen coverage in whole tissue libraries from patients with Barrett's 198
4.7.3 Boxplot of detected pathogen mapping rate in whole tissue libraries from patients with Barrett's 199
4.7.4 Pathogens detected frequently in single cell libraries from patients with Barrett's 201
4.7.5 Boxplot of detected pathogen coverage in single cell libraries from patients with Barrett's 202
4.7.6 Boxplot of detected pathogen mapping rate in single cell libraries from patients with Barrett's 203

A.1.1 BEARscc scores for various k - brain and blanks 222
A.1.2 BEARscc scores for various k - murine intestine 223
A.1.3 BEARscc scores for various k - murine brain 224
A.1.4 Histology of biopsies is verified for each patient and tissue 225
A.1.5 Histology of biopsies is verified for each patient and tissue 225
A.1.6 Barplot of inclusion/exclusion based on quality control filtering by patient and tissue 226
A.1.7 Barplot of inclusion/exclusion based on quality control filtering by batch and tissue 227
A.1.8 BEARscc metrics at chosen k for all-sample analysis cluster 6 . . 228
A.1.9 Barplot of tissue distribution across all-sample analysis clusters - all cells 229
A.1.10 Heatmap of BEARscc robust clusters of the esophagus 230
A.1.11 Barplot of patient distribution across esophageal clusters 231
A.1.12 Barplot of batch distribution across esophageal clusters 232
A.1.13 Scatterplots of esophageal BEARscc cell metrics 233
A.1.14 Heatmap of select genes differentially expressed in esophageal clusters235
A.1.15 Heatmap of genes differentially expressed in esophageal clusters . 236
A.1.16 Immunohistochemistry of marker genes in normal esophagus . . . 237
A.1.17 Gastric noise consensus heatmap 238
A.1.18 Barplot of patient distribution across gastric clusters 239
A.1.19 Barplot of batch distribution across gastric clusters 240
A.1.20 Scatterplots of gastric BEARscc cell metrics 241
A.1.21 Heatmap of select genes differentially expressed in gastric clusters 243
A.1.22 Heatmap of genes differentially expressed in gastric clusters . . . 244
A.1.23 Boxplot comparing select genes in gastric clusters 1 and 2 246
A.1.24 Duodenum noise consensus heatmap 247
A.1.25 Barplot of patient distribution across duodenal clusters 248
A.1.26 Barplot of batch distribution across duodenal clusters 249
A.1.27 Scatterplots of duodenal BEARscc cell metrics 250
A.1.28 Heatmap of select differentially expressed genes in duodenal clusters253
A.1.29 Heatmap of genes differentially expressed in duodenal clusters . . 254
A.1.30 Boxplot of LGR5 expression across intestinal clusters 255
A.1.31 Boxplot comparing select genes in duodenal cluster 5 256
A.1.32 Barrett's esophagus noise consensus heatmap 257
A.1.33 Barplot of patient distribution across Barrett's tissue clusters . . 258
A.1.34 Barplot of batch distribution across Barrett's tissue clusters . . . 259
A.1.35 Scatterplots of Barrett's tissue BEARscc cell metrics 260
A.1.36 Heatmap of select differentially expressed genes in Barrett's clusters263

A.1.37 Heatmap of genes differentially expressed in Barrett's clusters . . 264
A.1.38 Boxplot of LGR5 expression across Barrett's clusters 266
A.1.39 Heatmap of BEARscc robust clusters across cell types 267
A.1.40 Barplot of patient distribution across all-sample clusters 268
A.1.41 Barplot of batch distribution across all-sample clusters 269
A.1.42 Scatterplots of Barrett's tissue BEARscc cell metrics 270
A.1.43 Heatmap of genes differentially expressed in clusters from all-sample clustering 272
A.1.44 Heatmap of genes differentially expressed in clusters from all-sample analysis 273
A.1.45 Boxplot of immunoglobulin expression in cluster 1 by tissue . . . 274
A.1.46 Boxplot of total expression in libraries by BEARscc cluster 275
A.1.47 Boxplot of LGR5 expression across clusters of all-sample analysis 279
A.1.48 t-sne plots depicting metadata of single cells 280
A.1.49 Sankey diagram displays sample source and BEARscc cluster assignment 281
A.1.50 StemID analysis of esophageal and Barrett's cells 282
A.1.51 Immunohistochemistry of marker genes in Barrett's and duodenum 283
A.1.52 Immunofluorescence of goblet-type cells 284
A.1.53 Pathogens detected in single cell libraries from patients with Barrett's286

List of Tables

3.1 Samples sequenced by patient 83
3.2 Samples sequenced by batch 83
3.3 Samples sequenced by tissue 84
3.4 Cluster interpretation summary 145
3.5 Cluster interpretation summary - all cells 152

4.1 EBV detection in gastric cancer 187
4.2 *H. Pylori* detection in gastric cancer 188

A.1 Clinical and pathological data for selected patients 287
A.2 Samples sequenced by patient, tissue, and batch 288
A.3 BEARscc metrics for final clusters of individual tissues 288
A.4 Differential expression analysis exclusion summary - esophagus . . . 289
A.5 Differential expression analysis exclusion summary - gastric 289
A.6 Differential expression analysis exclusion summary - duodenum . . . 289
A.7 Differential expression analysis exclusion summary - Barrett's 290
A.8 Differential expression analysis exclusion summary - all samples . . . 290
A.9 Differential expression of esophageal single cell clusters 314
A.10 Differential expression of gastric single cell clusters 319
A.11 Differential expression of duodenal single cell clusters 327
A.12 Differential expression of Barrett's single cell clusters 333
A.13 Differential expression of clusters from the analysis of all QC-filtered cells 352
A.14 Differential expression analysis summary 353
A.15 BEARscc metrics for preliminary clusters in all-sample analysis . . . 353
A.16 BEARscc metrics for final clusters in all-sample analysis 354
A.17 MUC2, ITLN1 and SPINK4 expresssion in two sectinos from two BE patients 354

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

— Richard Feynman in *The Character of Physical Law*

1 Introduction

Contents

1.1 The problem in context **1**
1.2 Motivation **11**
1.3 Contribution and Scope **13**

1.1 The problem in context

Barrett's esophagus is a pre-malignant lesion leading to esophageal adenocarcinoma. The precise increase in risk for esophageal adenocarcinoma is difficult to discern, but previous work has shown that the lesion conveys risk ranging from 30 (Van der Veen *et al.* 1989) to more than 500 fold (Drewitz & Sampliner 1997). More recently, meta-data analysis has demonstrated that Barrett's esophagus conveys at least a 10 fold risk in presenting with esophageal adenocarcinoma (Solaymani-Dodaran 2004). From another perspective the incidence rate of Barrett's esophagus conversion into esophageal adenocarcinoma is between 0.2 and 2.1% (Theron *et al.* 2016; Holmberg *et al.* 2017; Hvid-Jensen *et al.* 2011). Approximately 1 in 20 people present with Barrett's esophagus (Zagari *et al.* 2016).

Barrett's esophagus occurs in the lower esophagus near the gastro-esophageal junction. In the gastro-intestinal tracts of healthy individuals, columnar epithelia

characterize the gastric and intestine tissues, while the normal esophagus is characterized by a stratified squamous epithelium (Takubo *et al.* 2008; Kim & Shivdasani 2016; Heath 1996). Broadly, Barrett's esophagus is defined by a phenomenon whereby the esophageal squamous epithelium is replaced by columnar features characteristic of those observed in the stomach and the duodenum (Booth & Thompson 2012), where the disease is thought to be a direct result of chronic gastric acid reflux (Savarino *et al.* 2016).

Importantly, columnar features present in the lower esophagus, typically stratified squamous epithelium in the healthy human, are sufficient for Barrett's diagnosis in Japan and the United Kingdom (Booth & Thompson 2012); however, while in the United States columnar epithelium is noted on endoscopy, only samples exhibiting intestinal metaplasias (e.g. the presence of goblet cells) are diagnosed with Barrett's esophagus (Booth & Thompson 2012). The disagreement on Barrett's diagnoses criteria coupled with imprecise guidelines on diagnosing dysplasia can confound attempts to characterize conversion rates from the pre-malignant Barrett's lesion into cancer (Vennalaganti *et al.* 2017). Extensive evidence suggests that risk for conversion to esophageal adenocarcinoma is restricted to columnar phenotypes with intestinal metaplasia (Booth & Thompson 2012). Therefore, in order to reduce ambiguity and maximize relevance to the conversion into cancer this work utilizes the United States guidelines for defining Barrett's esophagus. Because of the complex heterogeneity in Barrett's esophagus and the importance of that heterogeneity in risk and diagnosis, it is useful to briefly review the normal structure of the healthy esophageal, gastric, and duodenal tissue in greater detail before continuing.

The normal esophagus is composed of a simple flat squamous epithelium (Figure 1.1.1), which is most easily characterized by expression of specific keratins, but not keratinized (Takubo *et al.* 2008). Anatomically, the squamous epithelium of the esophagus in a healthy human is contiguous with the gastro-esophageal junction(Takubo *et al.* 2008). The most differentiated portion of the esophagus is found at the surface proximal to the lumen and consists of flat disk-like cells that are strongly positive for the keratin-4/keratin-13 (KRT-4/KRT-13) (KC *et al.*

2015; Ohashi *et al.* 2010). Moving toward the basement membrane, epithelial cells became less flat and less differentitated expressing keratin-5/keratin-14 (KRT-5/KRT-14)(KC *et al.* 2015; Ohashi *et al.* 2010) and commonly TP63 (Barbera *et al.* 2014). Approximately 10-20 layers of cells exist in the epithelium, of which approximately 15% consists of basal cells proximal to the basal membrane(Takubo *et al.* 2008). In addition to the flat squamous epithelium, the normal esophagus contains two gland-like structures, although these are reasonably rare in the normal esophagus (van Nieuwenhove *et al.* 2001). Mucosal glands are found in the mucosal layer of the esophagus and are restricted to the upper and lower sphincter of the esophagus (van Nieuwenhove *et al.* 2001). In contrast, esophageal submucosal glands are found throughout the length of the esophagus, and exist in the submucosa with excretory ducts crossing the mucosa and epithelium toward the esophageal lumen (Figure 1.1.2) (van Nieuwenhove *et al.* 2001; Long & Orlando 1999). The esophageal submucosal glandular ducts secrete products such as bicarbonates and mucins into the esophageal lumen, and potentially provide a protective role in humans and other mammals by base secretion mitigation of gastric acid exposure during reflux (Long & Orlando 1999).

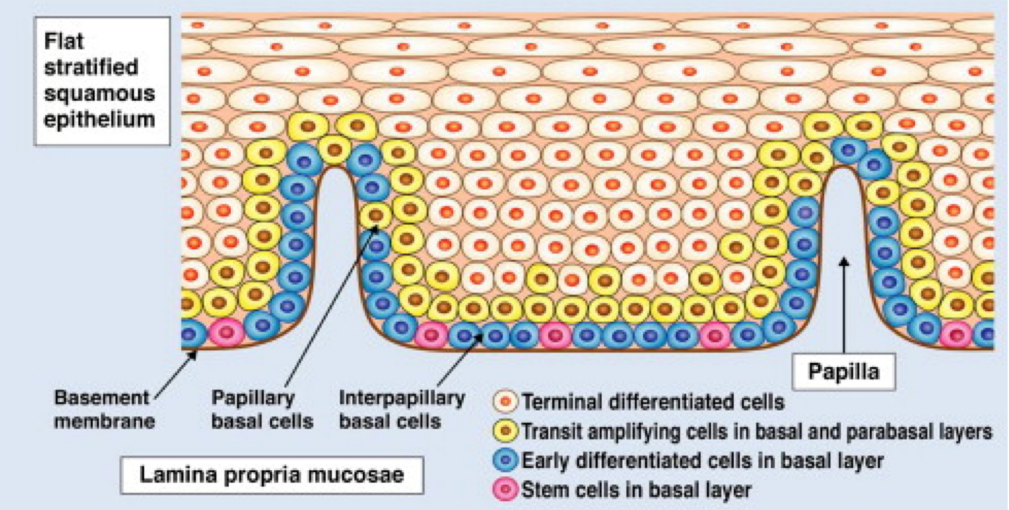


Figure 1.1.1: This figure and legend are directly reproduced from work done by Takubo *et al.* "Cellular organisation of the human oesophageal squamous epithelium. It is considered that stem and early differentiated cells are located in the basal layer on the basement membrane, transit amplifying cells are located in the parabasal layer and terminal differentiated cells are located above the parabasal layer (Takubo *et al.* 2008)."

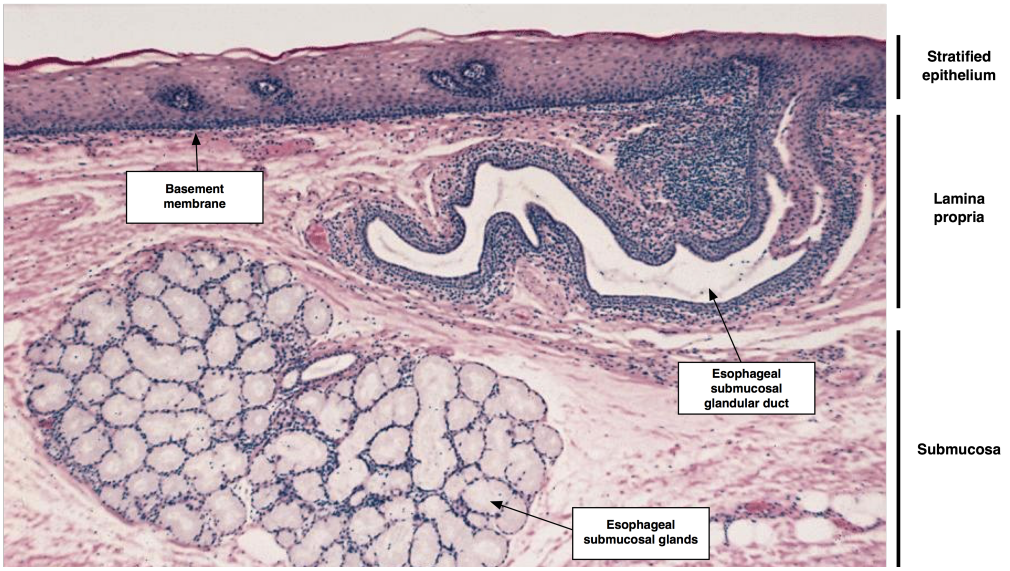


Figure 1.1.2: This figure and legend were adapted from work done by members of PathologyOutlines.com. "Section of esophageal mucosa shows a duct originating from the submucosal glands penetrating the squamous epithelium (PathologyOutlines.com 2013)." The submucosa, lamina propria and stratified squamous epithelium are designated with labels on the right of the figure. The esophageal submucosal glands and glandular ducts along with the basement membrane are indicated by labels and arrows within the histological slide.

In contrast to the esophagus, the cardia of the stomach is composed of a simple columnar epithelium organized into pits (Figure 1.1.3), which are the entryway into gastric glands (Hoffmann 2015; Bensley 1902). In the stomach, there are four anatomically defined regions: the cardia, fundus, corpus, and antrum (Hoffmann 2015; Bensley 1902). Depending on the anatomical location of the stomach, the proliferative cells that populate the stomach glands are thought to either reside in the bottom of the gland crypt or in the isthmus in the middle of the gland (Hoffmann 2015). The cardia region of the stomach, which is proximal to the gastro-esophageal junction, has glandular structure resembling those of the antrum (antral units), while the isthmus model for regeneration applies to fundic units found in the fundus and corpus (Hoffmann 2008; Hoffmann 2013). In general, the surface mucosa near the pits of the stomach are thought to express gastrokine-1 (GKN1) and gastrokine-2 (GKN2), while mucous neck cells are characterized by expression of pepsinogen (a zymogen) (PGC)(Hoffmann 2015). In the anatomical region nearest to the duodenum, the antrum (Bensley 1902), Clevers *et al* have demonstrated in elegant lineage tracing experiments that stem cell progenitors reside at the base of the crypt and are characterized by LGR5 expression (Barker *et al.* 2007; Barker *et al.* 2010). Due to the structural similarity between antral and cardiac glands, the LGR5 marked stem-cell observations in the antrum are likely the most pertinent in questions surrounding the origin of Barrett's esophagus. However, these stem cell identification results are largely derived from murine gastric tissue and may not directly correspond to human gastric glandular organization (Hoffmann 2015).

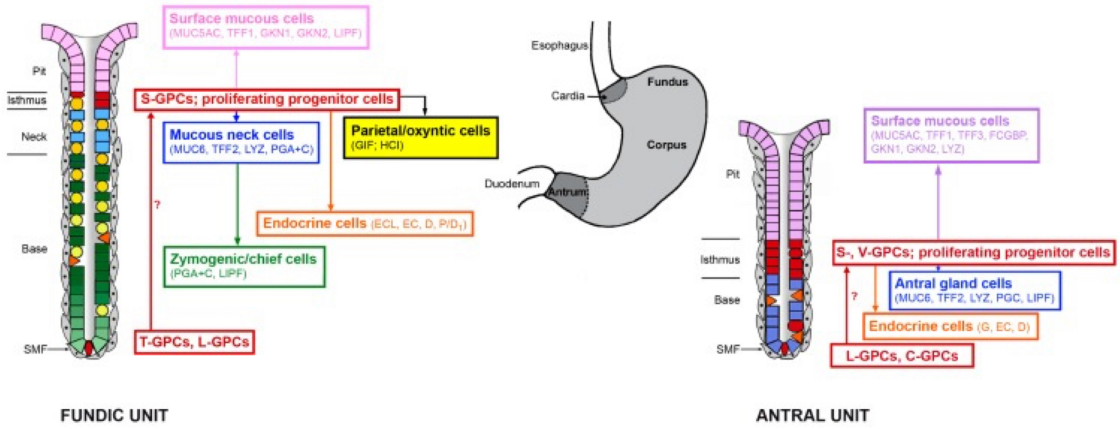


Figure 1.1.3: This figure and legend are directly reproduced from work done by Warner Hoffman (Hoffmann 2015). "Schematic representation of the two gross types of human gastric units and their continual, bidirectional self-renewal from gastric progenitor cells (GPCs)/stem cells. Shown are the major epithelial cell types and some of their characteristic secretory products such as MUC, mucins; TFF peptides, FCGBP, IgG Fc binding protein; GKN, gastrokines; LYZ, lysozyme; LIPF, gastric lipase, PGA and PGC, pepsinogens; GIF, intrinsic factor; and HCl, hydrochloric acid; as observed in the classical fundic and antral units, respectively. The various gastric precursor cell (GPCs)/stem cell populations are marked in red: C, CCK2 receptor+; L, Lgr5+; S, Sox2+; T, Troy+; V, villin+; EC(L), enterochromaffin(-like); G, gastrin; SMF, subepithelial myofibroblast; (modified and updated from [Hoffmann 2008; Hoffmann 2013]) (Hoffmann 2015)."

The duodenum is highly specialized relative to the gastric and esophageal epithelia (Figure 1.1.4). It is composed of a complex columnar organization containing cells specialized in both secretory and absorptive function (Carulli *et al.* 2014). In the duodenum, well-differentiated cells reside near the surface of crypts, while poorly differentiated proliferating cells and stem cells exist at the base of the crypt (Barker *et al.* 2007; Barker 2013; Grün *et al.* 2016b; Grün *et al.* 2015). Well-differentiated cells in intestinal crypts include secretory entero-endocrinocytes and goblet cells marked by chromogranin-A (CHGA) (Grün *et al.* 2015) and mucin-2 (MUC2) (Grün *et al.* 2015), respectively. Absorptive enterocytes are marked by intestinal alkaline phosphatase (ALPI) (Shin *et al.* 2014; Kim *et al.* 2014). Regulation of stem-cell maintenance and determinatino of absorptive vs secretory cell fate is understood to be accomplished primarily with NOTCH signaling (Pellegrinet *et al.* 2011; van Es *et al.* 2012; Kim *et al.* 2014). Notably, healthy

normal esophagus and gastric tissues are bereft of mucin-2 secreting goblet cells, but the Barrett's lesion is characterized by these goblet cells; an intriguing fact considering the goblet cell's anatomical niche is found in the distant small and large intestine in the healthy human.

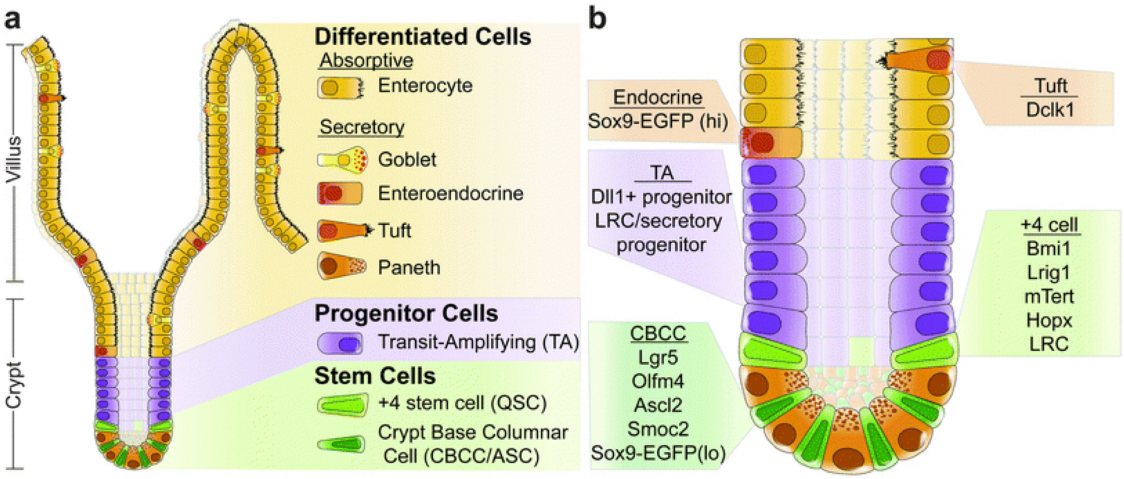


Figure 1.1.4: This figure and legend are directly reproduced from work by Carulli *et al* (Carulli *et al.* 2014). "Intestinal epithelial organization and markers. (a) The intestinal epithelium is organized into crypt and villus regions, with the stem and progenitor zone localized in the crypt. Current models favor the existence of two stem cell populations, the +4 stem cell and the crypt base columnar cell (CBCC), which are thought to be quiescent and active stem cells, respectively. Transit-amplifying (TA) progenitors arise from the stem cell compartment and differentiate into absorptive enterocytes or secretory goblet, enteroendocrine, tuft, or Paneth cells. Most of the differentiated cell populations migrate up the villi, but, uniquely, the Paneth cells move downward and reside between the CBCCs. (b) Molecular and functional markers that have been described for various proposed stem cell and potential stem cell populations. Of note, both TA cells and +4 cells have been shown to be Label Retaining Cells (LRCs). Sox9-EGFP has been shown to mark both CBCCs and clonogenic enteroendocrine cells, depending on the level of EGFP expression. The gene Dclk1 has been proposed to be a stem cell marker, but it has also been shown to be a specific marker of differentiated tuft cells. It is possible that there is an independent +4 cell population that is also marked with Dclk1, but this has not been verified by lineage tracing." (Carulli *et al.* 2014)

Anatomically, the lower esophagus is normally a stratified layer of squamous epithelium; in contrast Barrett's esophagus is a highly heterogenous, exhibiting the aforementioned features of the normal gastric and intestinal tissues (Gilbert *et al.* 2011; McDonald *et al.* 2015). The molecular definitions of these features at the cellular level are largely unknown. The tissue from which Barrett's esophagus

originates is an even greater mystery still. A number of diverse theories have been suggested over the last few decades with varying degrees of supporting experimental evidence (Hu *et al.* 2007; Kazumori 2006; Salo *et al.* 1996; Wang *et al.* 2011; Hutchinson *et al.* 2010; Quante *et al.* 2012; Jansen & Wright 2016; McDonald *et al.* 2014; Cancer Genome Atlas Research Network *et al.* 2017). However, the Barrett's origin debate continues, where a deeper understanding of the molecular determinants of the Barrett's phenotype could inform current hypotheses. Given the recent successes of the single cell transcriptomics in identifying subpopulations in heterogenous tissues as diverse as the murine brain (Zeisel *et al.* 2015), human melanoma (Tirosh *et al.* 2016), and identification of stem-like genes in single cell analysis of murine intestine (Grün *et al.* 2016b), it seems sensible to apply the technology to Barrett's esophagus in order to characterize the diverse histology in Barrett's at the molecular level.

In some ways the task of understanding the molecular determinants of Barrett's esophagus is ideally suited for single cell RNA-sequencing. The high heterogeneity present in the tissue coupled with its genomic stability and the absence of extensive necrotic tissue or white blood cell overwhelmed regions as compared to cancer mean that functional identification of sub-populations would likely be feasible with less cells than would be required in a fully developed cancer. The same features that make single cell RNA-sequencing analysis a suitable approach also diminish the likelihood of successful lineage tracing of Barrett's esophagus using DNA sequencing; the high genomic stability and low proliferation rates (relative to cancer) suggest that low clonality exists Barrett's esophagus. Low clonality makes detection of meaningful somatic variants for the purpose of lineage tracing with tools such as PyClone (Roth *et al.* 2014) unlikely. Therefore, single cell RNA-sequencing is a better approach for understanding the features behind the observed heterogeneity in Barrett's esophagus.

Unfortunately, single cell RNA-sequencing experiments are subject to extensive noise. In single cell sequencing experiments, the limited RNA concentration in a single cell leads to stochastic sampling effects that result in highly variable read

count measurements (Grün *et al.* 2014; Kim *et al.* 2015). Worse still, during library preparation there is a significant chance that expressed genes of interest will fail to be amplified even at moderately high transcript concentrations within a cell and no read counts will be observed (Grün *et al.* 2014; Kim *et al.* 2015). This false negative is commonly referred to as the drop-out effect (Grün *et al.* 2014; Kim *et al.* 2015). Aside from random technical variability, systematic variation due to subtle difference in cDNA conversion, library preparation and sequencing depth across batches are frequently observed (batch effects). Recent work by Hicks *et al* suggests that interpretation of numerous published results in single cell RNA-sequencing could be compromised by batch effects (Hicks *et al.* 2015). The problem of batch effects are further compounded by the logistical realities of single cell RNA-sequencing such as FACS sorting that often make simultaneous preparation of all factors of a variable of interest impossible; oftentimes single-cell RNA-sequencing experimental designs are necessarily confounded by batch due to logistics (Wagner *et al.* 2016). Finally, because single cell samples are necessarily destroyed during the experiment, technical replicates are an impossibility.

The high degree of technical variation is an important consideration when utilizing single cell RNA-sequencing to understand meaningfully the questions of origin and molecular determinants in Barrett's esophagus. Accurate identification and interpretation of cell sub-populations in Barrett's esophagus are necessary analytical steps when using a single-cell RNA sequencing approach for these questions. Therefore, characterizing the extent to which subpopulations identified in the experiment are robust to both systematic and random noise is essential. While single-cell RNA sequencing of Barrett's esophagus could illuminate the molecular defining features of Barrett's esophagus and thereby potentially provide new insights into its origin, factors extrinsic to the expression biology of patients could also be important in the etiology of the disease. Specifically, it is conceivable that the other aspects of the microenvironment of the esophagus in patients presenting with Barrett's could contribute to development of the disease.

Among the the most obvious external influences on human biology are viruses and microbes. In fact, pathogens have been implicated in cancer from the earliest days of the field of oncology (Hudnall 2014). Indeed, viruses are the primary mechanism for transformation of cells in order to establish immortalized cell lines (Ahuja *et al.* 2005; Gilbert 2001). The oncogenic risk associated with particular strains of the human papilloma virus (HPV) in cervical cancer is among the most well known examples of pathogen associated oncogenic risk (Faridi *et al.* 2011). An understanding of the importance of HPV in cervical cancer has led to vaccination efforts that have recently been shown to be successful in the reduction of cervical cancer incidence (Kim *et al.* 2016). In brain cancer, human herpesvirus 5, more commonly the cytomegalovirus, has been implicated (Söderberg-Nauclér 2015; Söderberg-Nauclér & Johnsen 2014; Wick & Platten 2014; Huse & Aldape 2014). In gastric cancer the bacterium *Helicobacter pylori* has been associated with malignancy (Peek & Blaser 2002; Ye *et al.* 2004; Aituov *et al.* 2012). Specifically, the CagA expressing variants confer an increased risk for developing gastric cancer to the extent that the World Health Organization has designated CagA as an oncoprotein (Peek & Blaser 2002; Ye *et al.* 2004). Human herpesvirus 4, more commonly Epstein-Barr virus (EBV), has been implicated in a sub-type of gastric cancer (Liu *et al.* 2015; Wang *et al.* 2014; The Cancer Genome Atlas Research Network 2014; Gulley 2015). EBV has also been frequently observed in nasopharyngeal squamous cell carcinoma (NPC) tumors and HPV is now understood to be important in a subset of oral squamous cell carcinomas (Reddout *et al.* 2007). Because pathogens have been implicated in numerous cancers of the aero-digestive tract, Barrett's esophagus could also have a microbial or viral component to its etiology. In fact, the question is of general interest to the esophageal adenocarcinoma research community; there is an ongoing clinical study investigating the role of the microbiome in gastro-esophageal reflux disease and progression to esophageal adenocarcinoma (Parsons 2011).

1.2 Motivation

Barrett's esophagus represents a significant risk for esophageal adenocarcinoma, where presentation with Barrett's esophagus implies a 10 fold increase in risk. Furthermore, 5-year survival for esophageal adenocarcinoma is extremely poor; only 5-15% of patients survive (Cunningham *et al.* 2010). Worse still, the incidence of esophageal adenocarcinoma has risen in recent decades (CRUK 2014; Cunningham *et al.* 2010). From these observations, it is not hard to conclude that esophageal adenocarcinoma is a disease of definitive need; any insights into the molecular mechanisms leading to the development of esophageal adenocarcinoma could have important clinical implications downstream in the forms of diagnostic, preventative, and drug-development initiatives. As a corollary, a deeper understanding of molecular definition and origins of the pre-malignant lesion, Barrett's esophagus, would be valuable clinically and in terms of translational research.

Understanding the origins and molecular features defining pre-malignant lesions, such as Barrett's esophagus, can provide insights into molecular mechanisms of progression leading to targets for drug discovery. In fact, in the wake of the Cancer Genome Atlas (TCGA) project (Weinstein *et al.* 2013; Tomczak *et al.* 2015), the lack of understanding of the molecular and cellular events defining the initiation and progression of cancer is a critical barrier in identifying new approaches for its prevention and detection (Campbell *et al.* 2016). Indeed, the American Association for Cancer Research recently called for a Pre-Cancer Genome Atlas (PCGA) (Campbell *et al.* 2016) to supplement the understanding of cancer gleaned from the TCGA project (Weinstein *et al.* 2013; Tomczak *et al.* 2015). An obvious example of the utility of understanding etiology of cancer is the importance of HPV in the onset of cervical cancer and relevant pre-malignant lesions (Kutlubay *et al.* 2013), which led to the implementation of preventative vaccines (Faridi *et al.* 2011). Should any such obvious host-pathogen implication exists in Barrett's esophagus, it would have important implications in cancer prevention.

As an example in Barrett's disease, Stachler *et al* utilized genomic studies of Barrett's esophagus in conjunction with fully formed esophageal adenocarcinoma

in order to refine our understanding of the mutational progression that occurs in Barrett's esophagus (Stachler *et al.* 2015). By understanding mutations important in esophageal adenocarcinoma progression through study of the pre-malignant lesion, targets for intervention may be identified. Similarly, by identifying the functional determinants behind the histological diversity in Barrett's esophagus, we may gain insights into the phenotypic switches important in moving from squamous to columnar histology. Furthermore, an understanding of these molecular determinants and their origin provides a theoretical framework in which to design experiments for identifying markers for progression. These markers could serve as new therapeutic targets to stall oncogenic progression or provide important information in terms of patient risk stratification.

Clinically, the characterization of pre-malignant lesions is important in stratifying patients for risk and decisions concerning early stage surveillance in cancer. Early detection in esophageal cancer is important in patient survival with 5 year survival for early stage detection at 83-90% compared to 10-15% survival in late stage cancers (Sharma 2005). Furthermore, esophageal adenocarcinoma detected due to surveillance of patients previously presenting with Barrett's esophagus are more likely to be early stage compared to those detected outside of surveillance (Sharma 2005). However, the cost-effectiveness of surveillance of Barrett's esophagus is debatable (Ballester & Cruz-Correa 2014). Moreover, screening and surveillance place physical (scoping procedures) and psychological (patient awareness of risk for cancer) burdens on patients. For the majority of patients that never progress to cancer, these burdens are unnecessary. To that end, identifying the tissue of origin in Barrett's esophagus and a relevant biomarker could facilitate earlier and risk-targeted screening of Barrett's esophagus. This would be particularly powerful when coupled with knowledge of molecular markers for Barrett's progression into cancer. Ultimately, an improved understanding of the functional determinants and origin of Barrett's could provide a framework to identify biomarkers for progression and refine screening and surveillance protocols using molecularly defined risk stratification.

This would simultaneously decrease the physical, psychological, and social costs of such protocols and increase the likelihood of detecting early stage adenocarcinomas.

1.3 Contribution and Scope

In order to understand the etiology of Barrett's esophagus, we sequenced the RNA of single cells from four patients with Barrett's esophagus and two patients with normal squamous epithelium. In the four patients that presented with Barrett's esophagus, endoscopic biopsy samples were collected from the Barrett's lesion itself. To control for the diverse histological features in Barrett's esophagus, endoscopic samples were also collected form the proximal normal esophagus, the stomach, and the duodenum. In two patients with a history of gastric reflux, no Barrett's was observed; for these patients, endoscopic biopsies were sampled from the normal squamous tissue only. Additionally, whole tissue RNA-sequencing was performed on a portion of all samples taken from patients with Barrett's disease. We hoped that by examining the cellular heterogeneity of the three control sites and Barrett's lesion together, we could observe a functional relationship providing deeper understanding of the intrinsic origins of Barrett's esophagus. In order to overcome the question of noise when examining inter-tissue relationships in these data, I developed BEARscc, a tool for examining clustering algorithm performance in the context of modeled noise.

In order to explore any influences of pathogen presence in the esophageal micro-environment in Barrett's esophagus, we determined to assay for pathogen and commensal RNA in our whole tissue and single-cell sequencing cohorts. I was unable to successfully implement PathSeq (Kostic *et al.* 2011) and the scope of tools such as VirusSeq (Chen *et al.* 2013), VY-PER (Forster *et al.* 2015) and similar software is restricted to viral analysis; therefore, I elected to develop my own tool, poshTitan, for detection of pathogens and commensals in next-generation sequencing data. We aimed to use poshTitan to refute or provide evidence for the hypothesis that pathogens are associated with Barrett's esophagus.

Here I describe two novel computational methods designed to enhance analysis of single cell RNA-sequencing data: BEARscc (Chapter 3) and poshTitan (Chapter 5). BEARscc is available as an R package, and can be found at:

https://bitbucket.org/bsblabludwig/bearscc.

In contrast, poshTitan requires further work before widespread dissemination. The source code in its current state is available at:

https://bitbucket.org/bsblabludwig/poshtitan.

Both algorithms have been applied to the previously described cohort of single cell libraries sequenced from endoscopic biopsies of four patients' proximal normal esophagus, duodenum, stomach, and Barrett's lesions and from endoscopic biopsies of normal squamous esophagus from two patients without Barrett's esophagus in order to provide further insight into the etiology of Barrett's esophagus.

BEARscc, or Bayesian ERCC Assessment of Robustness of single-cell clusters, models technical noise in single cell RNA-sequencing experiments from ERCC spike-in controls (Baker *et al.* 2005) and uses the model to create simulated technical replicates (Chapter 3). These simulated technical replicates are then used to evaluate clusters on single-cell expression data by generating a noise consensus across technical replicates. BEARscc also includes metrics to assist the user in interpreting the noise consensus. We demonstrate that BEARscc accurately models real technical noise and illustrate its effectiveness in controlled clustering experiments. We also demonstrate the utility of BEARscc in providing biological clarity by applying the algorithm to two publicly available single cell RNA-sequencing datasets.

Furthermore, BEARscc has been applied to our cohort of upper GI single cell libraries obtained from six patients (Chapter 4). BEARscc successfully revealed robust relationships between cells of the upper GI tract. Specifically, BEARscc verified profound functional similarity between Barrett's cells and esophageal submucosal glandular ducts. The algorithm also defined robust groups of known cell types in the esophagus, the stomach, and the duodenum. Analysis downstream of BEARscc identified stem-like marker, OLFM4, as expressed in Barrett's cells with high similarity to esophageal submucosal glandular ducts. A developmental marker,

LEFTY1, was also identified for the same subpopulation of cells. Additionally, two new markers, ITLN1 and SPINK4, were identified using the subpopulations identified by BEARscc.

The poshTitan pipeline identifies pathogen and commensal reads using Kraken (Wood & Salzberg 2014) and Bowtie2 (Langmead & Salzberg 2012) software (Chapter 5). Specifically, the poshTitan first uses the Kraken to assign reads to potential pathogens and commensals, and then aligns flagged reads to their respective genomes with Bowtie2. From the Bowtie2 alignments, the poshTitan produces summary statistics such as coverage and mapping data for each putative pathogen in order to make a binary call on the presence or absence of a foreign genome. Applying the poshTitan to whole tissue and single-cell libraries from upper GI endoscopic biopsies supported the prevailing idea that pathogens and commensals are unimportant in Barrett's esophagus.

Though this be madness, yet there is method in 't.

— Polonius, conversing with Hamlet in William Shakespeare's *Hamlet*

All you need in this life is ignorance and confidence, and then Success is sure.

— Mark Twain, *Notebook, 1887*

2

BEARscc models noise to reveal cluster robustness

Contents

2.1 Introduction **18**
2.2 Data and methods **21**
2.2.1 Data 21
2.2.2 Clustering of counts data 22
2.2.3 Estimation of cluster number k 22
2.2.4 Gene sampling for benchmarking 23
2.3 Algorithm description and implementation **23**
2.3.1 BEARscc workflow 23
2.3.2 Algorithmic generation of simulated technical replicates . 26
2.3.3 Modelling noise from spike-ins 27
2.3.4 Inference of transcript dropout distributions using spike-ins 29
2.3.5 Computation of consensus matrix 31
2.3.6 Computation of BEARscc cluster metrics 32
2.3.7 Computation of BEARscc cell metrics 32
2.3.8 Implementation 32
2.4 Optimizing BEARscc to accurately simulate technical replicates **33**
2.4.1 Pseudocode illustrating the variations in implementation for the BEARscc algorithm 35
2.4.2 Gene-wise implementations of BEARscc are superior to global implementations 36
2.4.3 A single drop-out function most accurately models observed drop-out number 43
2.5 BEARscc accurately models technical noise in single cell RNA-seq data **46**
2.6 The importance of normalization in clustering analysis **48**

2.7 Performance of BEARscc combined with different clustering algrotihms **55**
2.8 Choosing the number of simulated replicates **60**
2.9 BEARscc identifies cells and clusters that are robust to noise . . **62**
2.9.1 An illustration in murine intestine single cell samples . . 62
2.9.2 An illustration in murine brain single cell samples 67
2.10 Benchmarking BEARscc against simple gene sampling **70**
2.11 Conclusions and discussion **72**

2.1 Introduction

Single cell messenger RNA sequencing (scRNA-seq) has recently been shown to effectively characterize cellular heterogeneity of complex eukaryotic tissues as diverse as the mammalian brain (Zeisel *et al.* 2015; Darmanis *et al.* 2015), blood (Mahata *et al.* 2014), and intestine (Grün *et al.* 2015). The technique can be used to establish an atlas of functional subtypes of cells in normal tissue (Grün *et al.* 2015), and to identify micro-populations of cells relevant to diseases such as cancer (Tirosh *et al.* 2016). In many scRNA-seq studies, statistical clustering methods are used to identify cells with similar gene expression profiles that could represent distinct cell types (Grün *et al.* 2016b; Zeisel *et al.* 2015; Grün *et al.* 2015).

However, technical variability poses a major hurdle in the analysis of scRNA-seq data (Grün *et al.* 2014; Kim *et al.* 2015). Two primary forms of random experimental noise confound single cell analyses. Firstly, read count measurements vary considerably due to the limited amount of starting material, leading to stochastic sampling effects (Grün *et al.* 2014; Kim *et al.* 2015). This problem is exacerbated with decreasing transcript length and concentration. Secondly, even when there are moderately high transcript concentrations within the cell, there is a significant chance that an expressed transcript of interest will not be amplified in library preparation. In this case, no reads for the transcript are observed in the experiment (the *drop-out* effect), resulting in a false-negative (Grün *et al.* 2014; Kim *et al.* 2015).

In addition to substantial random technical variability, systematic variation due to sample processing can confound clustering results. For logistical reasons, single cell samples are frequently prepared in multiple batches. Batch-dependent differences in cDNA conversion, library preparation and sequencing depth can easily mask biological differences between cells. For example, Hicks *et al.* recently reported that batch differences can skew the relationship between drop-out effects and median expression of detected genes (Hicks *et al.* 2015). Worryingly, Hicks *et al.* claim that batch effects might compromise the interpretation of many published scRNA-seq results (Hicks *et al.* 2015).

One approach to measure technical variation between samples that has become widely adopted is the addition of known quantities of RNA to each cell sample before cDNA conversion and library preparation. The widely used External RNA Control Consortium (ERCC) library (Jiang *et al.* 2011) consists of 92 bacterial transcripts of varying GC content, length, and concentration (Vallejos *et al.* 2015) is an example of one such set of know quantities of RNA. Several methods have been developed that make use of spike-in controls (like the ERCC spike-ins) to normalize read counts per cell before further analysis, in order to reduce the influence of batch and sample differences (Vallejos *et al.* 2015; Brennecke *et al.* 2013). However, the use of ERCC spike-ins as a tool for normalizing read counts has recently been criticized for exacerbating the influence of differences in cell size and RNA content (Wagner *et al.* 2016; Vallejos *et al.* 2015). Instead of using spike-ins for normalization of read counts, I hypothesized that ERCC spike-ins could be used to measure the distribution of experimental technical variation across samples. Crucially, this approach would not rely on the assumption that spike-ins are perfectly correlated with the absolute abundance of endogenous transcripts per cell. The resulting model of technical noise could then be used to simulate technical replicates and thus directly evaluate the sensitivity of clustering results to technical variation, which has not previously been possible.

The list of clustering methods designed for single cell data is extensive and growing (Grün & van Oudenaarden 2015; Bacher & Kendziorski 2016; Wagner

et al. 2016). For the purpose of this analysis, I focused on three recently developed algorithms that have been used to analyse cellular composition of tissues in a number of highly influential recent publications: RaceID2 (Grün *et al.* 2015; Grün *et al.* 2016b), BackSPIN (Zeisel *et al.* 2015), and SC3 (Kiselev *et al.* 2016). RaceID2 (Grün *et al.* 2015; Grün *et al.* 2016b) uses the k-medoids algorithm to cluster libraries and identify outlier cells. BackSPIN (Zeisel *et al.* 2015) relies on the SPIN algorithm that iteratively splits the correlation matrix, assigning genes to the groups of cells formed by each split. Finally, SC3 (Kiselev *et al.* 2016) combines three clustering metrics with two dimensionality reduction methods followed by iterative k-means clustering to establish a consensus across many algorithmic solutions to the clustering problem. All three algorithms take different approaches to predicting related groups of cells, but none of them provides a measure of confidence in the robustness of the detected clusters.

Here, I present BEARscc, a tool to evaluate cluster robustness in the face of technical variability (Section 2.3). BEARscc models technical noise from RNA spike-in controls, most commonly ERCC spike-in controls, and utilizes the resulting model to create simulated technical replicates. Each simulated replicate is reclustered and a consensus clustering, the *noise consensus matrix*, is formed across simulated technical replicates. First I explain the data and computational methods utilized throughout the chapter (Section 2.2) . Next, I explain how I refined the simulation of technical replicates from several potential implementations of BEARscc (Section 2.4), and I demonstrate that the finalized implementation of simulating technical replicates accurately models experimental noise from ERCC spike-in measurements (Section 2.5). I show that BEARscc is compatible with three prominent clustering algorithms and I highlight the importance of normalization in relation to batch effects in the context of these three clustering algorithms (Section 2.6). I demonstrate that BEARscc in conjunction with three prominent clustering methods clearly identifies artifactual clusters that are a result of systematic and random technical variation (Section 2.7). I assess the number of simulated technical replicates necessary to obtain a stable noise consensus matrix (Section 2.8). Finally, using

two publicly available datasets as an example, I describe how BEARscc analysis improves the interpretation of clustering results and provides additional biological insights (Section 2.9).

2.2 Data and methods

The following describes in detail the relevant data and methods employed throughout this chapter.

2.2.1 Data

Brain whole tissue technical replicates

Brain whole tissue total RNA (Agilent Technologies, cat 540005) was diluted to 10pg aliquots and added to a $1\mu L$ cDNA master mix solution containing ERCC spike-in (Baker *et al.* 2005) solution diluted 1:10000. Sample preparation, cDNA conversion, library preparation, and sequencing were performed by the Wellcome Trust Center for Human Genomics Sequencing Core. Samples were pipetted into 96-well plates and treated as single cells using Smartseq2 cDNA conversion as described by Picelli *et al* (Picelli *et al.* 2013) with minor modifications. The library was prepared using Fludigm's recommendations for Illumina NexteraXT at $\frac{1}{4}$ volume with minor modifcations, and sequenced on the Illumina HiSeq4000 platform. Raw reads were mapped to hg19 with STAR (Dobin *et al.* 2013). Exact position duplicates were filtered, and fragment counts for genes were computed with HTseq (Anders *et al.* 2015).

Blank whole tissue technical replicates

DNase free water was added to $1\mu L$ cDNA master mix solution containing ERCC spike-in (Baker *et al.* 2005) solution diluted 1:10000. Sample preparation, cDNA conversion, library preparation, and sequencing were performed by the Wellcome Trust Center for Human Genomics Sequencing Core. Samples were pipetted into 96-well plates and treated as single cells using Smartseq2 cDNA conversion as described by Picelli et al (Picelli *et al.* 2013) with minor modifications. The

library was prepared using Fludigm's recommendations for Illumina NexteraXT at $\frac{1}{4}$ volume with minor modifcations, and sequenced on the Illumina HiSeq4000 platform. Raw reads were mapped to hg19 with STAR (Dobin *et al.* 2013). Exact position duplicates were filtered, and fragment counts for genes were computed with HTseq (Anders *et al.* 2015).

Publicly available murine intestinal, hippocampus, and cortex single cell data

Primary murine cortex and hippocampus single cells measurements (3005 cells from Zeisel et al.(Zeisel *et al.* 2015)) were used. ERCC and sample count data were retrieved from the publicly available Linnarsson laboratory data repository (http://linnarssonlab.org/cortex/). Primary murine intestinal single cell measurements (291 cells from Grün et al (Grün *et al.* 2016b).) were used. ERCC and sample count data were retrieved form publicly available Van Oudenaarden github repository. (https://github.com/dgrun/StemID).

2.2.2 Clustering of counts data

BackSPIN (Zeisel *et al.* 2015), SC3 (Kiselev *et al.* 2016) and RaceID2 (Grün *et al.* 2016b) were run according to algorithm-specific recommendations. RaceID2 was allowed to identify cluster number under default parameters. For the brain and blanks control experiment data, RaceID2 was modified to skip normalization since scaled counts per million normalization had already been applied to the data set. The number of clusters, k, selected for SC3 clustering was determined empirically by selecting k with the optimal silhouette distribution across noise injected counts matrices.

2.2.3 Estimation of cluster number k

In order to determine the cluster number, k, from the hierarchical clustering of the noise consensus, the resulting dendrogram was cut multiple times to form N clusterings with cluster numbers k=1 to k=N clusters. The average score metric was computed for each clustering, and k was chosen by taking the k with the

maximum average score metric. Evaluating all possible k from 1 to the number of cells in the experiment is computationally expensive and unlikely to be biologically meaningful. In this work, N was capped at 0.1 times the number of cells in the experiment: N=10 for the brain and blanks control, N=30 for the murine intestine experiment, and N=300 for the murine brain data.

2.2.4 Gene sampling for benchmarking

For comparison with BEARscc, 100 subsampling iteration matrices for n cells and m genes were generated by sampling one half of expressed genes and clustered using the respective clustering algorithm (SC3, BackSPIN, RaceID2). For each dataset, genes were excluded with less than 25 total raw counts across all samples in the cohort. The remaining genes formed the sample space. In each subsampling iteration, one half of the genes were sampled without replacement, and their expression across cells was used as the counts matrix. Identically to the computation of the BEARscc noise consensus matrix, cluster labels were used to compute an n x n binary association matrix for each clustering, and an arithmetic mean was taken for each respective element across the resulting 100 association matrices to produce an n x n subsampling consensus matrix. Identically to BEARscc analysis, the BEARscc score metric was used to determine cluster number k, and the resulting cluster labels for each dataset and algorithm were compared with BEARscc by computing the adjusted rand index for each with respect to the relevant ground truth.

2.3 Algorithm description and implementation

2.3.1 BEARscc workflow

In order to simulate technical replicates, BEARscc first builds a statistical model of expression variance and drop-out frequency, which is dependent only on observed gene expression. The parameters of this model are estimated from spike-in counts (Figure 2.3.1, Step 1). Expression-dependent variance is approximated by fitting read counts of each spike-in transcript across cells to a mixture model comprised of a Poisson and negative binomial distribution (Section 2.3.3). The drop-out

model (Section 2.3.4) in BEARscc has two distinct parts: the *drop-out injection distribution* models the likelihood that a given transcript concentration will result in a drop-out, and the *drop-out recovery distribution* models the likelihood that an observed drop-out resulted from a given transcript concentration. The drop-out injection distribution is taken to be the observed drop-out rate in spike-in controls as a function of actual spike-in transcript concentration. This distribution is then used to estimate the drop-out recovery distribution density via Bayes' theorem and a an empirically informed set of priors and assumptions. Briefly, BEARscc utilizes the drop-out injection distribution and the number of observed zeroes for each endogenous gene to infer a gene-specific probability distribution describing the likelihood that an observed drop-out should in fact have been some non-zero value, given the drop-out rate of the endogenous gene.

In the second step, BEARscc generates simulated technical replicates by applying the models described in Step 1 (Figure 2.3.1, step 2 and Section 2.3.2). For every observed count in the range of values where drop-outs occurred amongst the spike-in transcripts, BEARscc uses the drop-out injection distribution from Step 1 to determine whether to convert the count to zero. For observations where the count is zero, the drop-out recovery distribution is used to estimate a new value, given the overall drop-out frequency for the gene (Section 2.3.2). Next, BEARscc substitutes all values larger than zero with a value generated from the derived model of expression variability, parameterized to the observed count for that gene. This procedure can then be repeated any number of times to generate a collection of simulated technical replicates.

In the third step, the simulated technical replicates are then re-clustered, using exactly the same method as for the observed data; this re-clustering for each simulated technical replicate is described as an *association matrix* where each element indicates whether two cells share a cluster identity (1) or cluster apart from each other (0) (Figure 2.3.1, step 3). The association matrices for each simulated technical replicate are averaged to form a noise consensus matrix that can be easily interpreted (Section 2.3.5). Each element of the noise consensus matrix represents

the fraction of simulated technical replicates that, upon applying the clustering method of choice, resulted in two cells clustering together (the *association frequency*).

To improve the utility of the resulting noise consensus matrix, I provide several downstream analytical tools with BEARscc. I developed three cluster metrics (*stability*, *promiscuity* and *score*, see Section 2.3.6) that can be used to qualitatively analyse a given clustering result. All three metrics can also be calculated on a per-cell basis (Section 2.3.7). In that case, *stability* (and *promiscuity*) denote the frequency with which a single cell sample is clustered within (or outside of) its original cluster. In addition to these summary statistics, BEARscc can also derive a set of "consensus clusters" from the noise consensus matrix by applying complete-linkage clustering. To determine the optimal number of clusters, k, in the consensus cluster set, I propose to use the k that gives a maximum average cluster score, where k values of 1 to a manually chosen upper boundary of cluster number N are examined (Section 2.2.3) . Intuitively, this definition results in the fewest possible clusters that maximise internal consistency within clusters while minimizing changes between clusters across simulated technical replicates. Together these tools provide an intuitive way to describe and interpret single-cell transcriptome experiments. I suggest users define N to be the number of cells in the cohort divided by 10. However, biological and technical aspects of the experiment may inform the limit the maximum cluster number N examined.

An illustration of the workflow of BEARscc algorithm

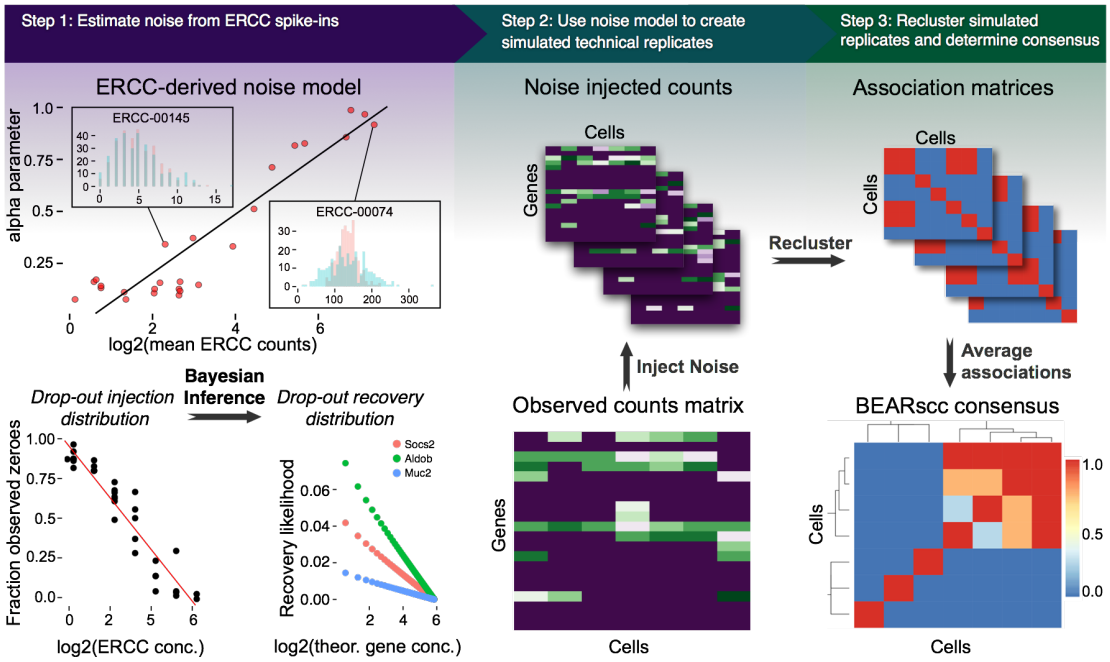


Figure 2.3.1: BEARscc algorithm overview. **Step 1**, Learning the noise model: the distribution of observed ERCC reads changes with their abundance (blue in insets) and each ERCC is described using a mixture model with parameter α (red in insets). Linear regression over all ERCC-derived α values is then used to estimate α for any endogenous gene as a function of mean expression (top left). Dropout effects are modeled separately. The drop-out injection distribution (bottom far left) is modeled based on the frequency of zero counts in ERCCs at known concentration. The drop-out recovery distribution for each gene is then determined via Bayesian inference from this distribution. This distribution determines the probability that a drop-out is permuted to a value larger than 0 (shown for 3 example genes on bottom left). **Step 2**, simulating replicates: an observed counts matrix (centre bottom) is transformed into multiple simulated technical replicates by applying the noise model derived in Step 1. **Step 3**, calculating a consensus: concordance matrices that result from clustering each noise injected counts matrix (top right) are combined into a single consensus matrix that reflects the frequency of with which cells are observed in the same cluster across all simulated replicates.

2.3.2 Algorithmic generation of simulated technical replicates

Simulated technical replicates were generated from the noise mixture-model and two drop-out models. For each gene, the count value of each sample is systematically transformed using the mixture-model, $Z(c)$, and the drop-out injection, $\Pr(X = 0 \mid Y = y)$, and recovery, $\Pr(Y_j = y \mid X_j = 0)$, distributions in order to generate simulated technical replicates as indicated by the following pseudocode:

FOR EACH gene, j
  FOR EACH count, c
    IF $c = 0$
      $n \leftarrow$ SAMPLE one count, y, from $\Pr(Y_j = y \mid X_j = 0)$
      IF $n = 0$
        $c \leftarrow 0$
      ELSE
        $c \leftarrow$ SAMPLE one count from $Z(n)$
      ENDIF
    ELSE
      IF $c \leq k$
        $dropout \leftarrow TRUE$ with probability, $\Pr(X = 0 \mid Y = k)$
        IF $dropout = TRUE$
          $c \leftarrow 0$
        ELSE
          $c \leftarrow$ SAMPLE one count from $Z(c)$
        ENDIF
      ELSE
        $c \leftarrow$ SAMPLE one count from $Z(c)$
      ENDIF
    ENDIF
    RETURN c
  DONE
DONE

2.3.3 Modelling noise from spike-ins

Technical variance was modeled by fitting a single parameter mixture model, $Z(c)$, to the spike-ins' observed count distributions. The noise model was fit independently

for each spike-in transcript and subsequently regressed onto spike-in mean expression to define a generalized noise model. This was accomplished in three steps:

1. Define a mixture model composed of *poisson* and *negative binomial* random variables:

$$Z \sim (1-\alpha) * Pois(\mu) + \alpha * NBin(\mu, \sigma) \tag{2.1}$$

2. Empirically fit the parameter, α_i, in a spike-in specific mixture-model, Z_i, to the observed distribution of counts for each ERCC spike-in transcript, i, where μ_i and σ_i are the observed mean and variance of the given spike-in. The parameter, α_i, was chosen such that the error between the observed and mixture-model was minimized.

3. Generalize the mixture-model by regressing α_i parameters and the observed variance, σ_i, onto the observed spike-in mean expression, μ_i. Thus the mixture model describing the noise observed in ERCC transcripts was defined solely by μ, which was treated as the count transformation parameter, c, in the generation of simulated technical replicates .

In step 2, a mixture model distribution is defined for each spike-in, i: $Z_i(\alpha_i, \mu_i, \sigma_i) \sim (1-\alpha_i) * Pois(\mu_i) + \alpha_i * NBin(\mu_i, \sigma_i)$. The distribution, Z_i, is fit to the observed counts of the respective spike-in, where α_i is an empirically fitted parameter, such that the α_i minimizes the difference between the observed count distribution of the spike-in and the respective fitted model, Z_i. Specifically, for each spike-in transcript, μ_i and σ_i were taken to be the mean and standard deviation, respectively, of the observed counts for spike-in transcript, i. Then, α_i was computed by empirical parameter optimization; α_i was taken to be the $\alpha_{i,j}$ in the mixture-model, $Z_{i,j}(\alpha_{i,j}, \mu_i, \sigma_i) \sim (1-\alpha_{i,j}) * Pois(\mu_i) + \alpha_{i,j} * NBin(\mu_i, \sigma_i)$, found to have the least absolute total difference between the observed count density and the density of the fitted model, Z_i. In the case of ties, the minimum $\alpha_{i,j}$ was chosen.

In step 3, $\alpha(c)$ was then defined with a linear fit, $\alpha_i = \alpha * log2(\mu_i) + b$. $\sigma(c)$ was similarly defined, $log2(\sigma_i) = \alpha * log2(\mu_i) + b$. In this way, the observed distribution

of counts in spike-in transcripts defined the single parameter mixture-model, $Z(c)$, used to transform counts during generation of simulated technical replicates:

$$Z(c) \sim (1 - \alpha(c)) * Pois(c) + \alpha(c) * NBin(c, \sigma(c)) \tag{2.2}$$

During technical replicate simulation, the parameter c was set to the observed count value, a, and the transformed count in the simulated replicate was determined by sampling a single value from $Z(c = a)$.

2.3.4 Inference of transcript dropout distributions using spike-ins

A model of the drop-outs was developed in order to inform the permutation of zeros during noise injection. The observed zeros in spike-in transcripts as a function of actual transcript concentration and Bayes' theorem were used to define two models: the *drop-out injection distribution* and the *drop-out recovery distribution*.

The drop-out injection distribution was described by $\Pr(X = 0 \mid Y = y)$, where X is the distribution of observed counts and Y is the distribution of actual transcript counts; the density was computed by regressing the fraction of zeros observed in each sample, D_i, for a given spike-in, i, onto the expected number spike-in molecules in the sample, y_i, e.g. $D = a * y + b$. Then, D describes the density of zero-observations conditioned on actual transcript number, y, or $\Pr(X = 0 \mid Y = y)$. Notably, each gene was treated with an identical density distribution for drop-out injection.

In contrast, the density of the drop-out recovery distribution, $\Pr(Y_j = y \mid X_j = 0)$, is specific to each gene, j, where X_j is the distribution of the observed counts and Y_j is the distribution of actual transcript counts for a given gene. The gene-specific drop-out recovery distribution was inferred from drop-out injection distribution using Bayes' theorem and a prior. This was accomplished in 3 steps:

1. For the purpose of applying Bayes' theorem, the gene-specific distribution, $\Pr(X_j = 0 \mid Y_j = y)$, was taken to be the the drop-out injection density for all genes, j.

2. The probability that a specific transcript count was present in the sample, $\Pr(Y_j = y)$, was a necessary, but empirically unknowable prior. Therefore, the prior was defined using the law of total probability, an assumption of uniformity, and the probability that a zero was observed in a given gene, $\Pr(X_j = 0)$. The probability, $\Pr(X_j = 0)$, was taken to be the fraction of observations that were zero for a given gene. This was done in order to better inform the density estimation of the gene-specific drop-out recovery distribution.

3. The drop-out recovery distribution density was then computed by applying Bayes' theorem:

$$\Pr(Y_j = y \mid X_j = 0) = \frac{\Pr(X_j = 0 \mid Y_j = y) \cdot \Pr(Y_j = y)}{\Pr(X_j = 0)}, \tag{2.3}$$

In the second step, the law of total probability, an assumption of uniformity, and the fraction of zero observations in a given gene were leveraged to define the prior, $\Pr(Y_j = y)$. First, a threshold of expected number of transcripts, k in Y, was chosen such that k was the maximum value for which the drop-out injection density was non-zero. Next, uniformity was assumed for all expected number of transcript values, y greater than zero and less than or equal to k; that is $\Pr(Y_j = y)$ was defined to be some constant probability, n. Furthermore, $\Pr(Y_j = y)$ was defined to be 0 for all $y > k$. In order to inform $\Pr(Y_j = y)$ empirically, $\Pr(Y_j = 0)$ and n were derived by imposing the law of total probability (Equation 2.4) and unity (Equation 2.5) yielding a system of equations:

$$\Pr(X_j = 0) = \Sigma_{y=0}^{k-1}(\Pr(X_j = 0 \mid Y_j = y) \cdot \Pr(Y_j = y)) \tag{2.4}$$

$$\Sigma_{y=0}^{k-1}\Pr(Y_j = y) = \Pr(Y_j = 0) + (k-1) \cdot n = 1 \tag{2.5}$$

The probability that a zero is observed given there are no transcripts in the sample, $\Pr(X_j = 0 \mid Y_j = 0)$, was assumed to be 1. With the preceding assumption, solving for $\Pr(Y_j = 0)$ and n gives:

$$n = \frac{1 - \Pr(Y_j = 0)}{k - 1} \tag{2.6}$$

$$\Pr(Y_j = 0) = \frac{\Pr(X_j = 0) - \frac{1}{k-1} \cdot \Sigma_{y=1}^{k-1}(\Pr(X_j = 0 \mid Y_j = y))}{(1 - \frac{1}{k-1} \cdot \Sigma_{y=1}^{k-1}(\Pr(X_j = 0 \mid Y_j = y))} \tag{2.7}$$

In this way, $\Pr(Y_j = 0)$ was defined by (Equation 2.6) for y in Y_j less than or equal to k and greater than zero, and defined by (Equation 2.7) for y in Y_j equal to zero. For y in Y_j greater than k, the prior $\Pr(Y_j = y)$ was defined to be equal to zero.

In the third step, the previously computed prior, $\Pr(Y_j = y)$, the fraction of zero observations in a given gene, $\Pr(X_j = 0)$, and the drop-out injection distribution, $\Pr(X_j = 0 \mid Y_j = y)$, were utilized to estimate with Bayes? theorem the density of the drop-out recovery distribution, $\Pr(Y_j = y \mid X_j = 0)$. During the generation of simulated technical replicates for zero observations and count observations less than or equal to k, values were sampled from the drop-out recovery and injection distributions as described in the pseudocode of the algorithm.

2.3.5 Computation of consensus matrix

100 simulated replicate matrices for n cells and m genes were clustered using the respective clustering algorithm (SC3, BackSPIN, RaceID2) as described above. Cluster labels were used to compute an n x n binary association matrix for each clustering. Each element of the association matrix represents a cell-cell interaction, where a value of 1 indicates that two cells share a cluster and a value of 0 indicates two cells do not share a cluster. An arithmetic mean was taken for each respective element across the resulting 100 association matrices to produce an n x n noise consensus matrix, where each element represents the fraction of noise injected counts matrices that, upon clustering, resulted in two cells sharing a cluster.

2.3.6 Computation of BEARscc cluster metrics

To calculate cluster *stability*, the noise consensus matrix was subset to cells assigned to the cluster. The cluster *stability* was then calculated as the arithmetic mean of the upper triangle of the subset noise consensus matrix. To calculate cluster *promiscuity*, the rows of the noise consensus matrix were subset to cells assigned to the cluster and the columns are subset to the cells not assigned to the cluster. For clusters with as many or more cells assigned to them than not assigned, the *promiscuity* was defined as the arithmetic mean of the elements in the subset matrix. Otherwise, the columns were further subset to the same number of cells as were assigned to the cluster, where the cells outside of the cluster with the strongest mean association with cells inside the cluster are chosen. The *promiscuity* was defined as the arithmetic mean of the elements in this further subset matrix. Each cluster's *promiscuity* was subtracted from its *stability* to calculate cluster *score*.

2.3.7 Computation of BEARscc cell metrics

To calculate a cell's *stability,* the arithmetic mean was taken of that cell's association frequencies with other cell's within the cluster. To calculate a cell's *promiscuity*, there were two cases. For cells in clusters with as many or more cells assigned to them than not assigned, the *promiscuity* was the arithmetic mean of that cell's association frequencies with all cells not assigned to the relevant cluster. For cells in clusters of size n, with fewer cells assigned to them than not assigned, the cell's *promiscuity* was the arithmetic mean with the n cells not assigned to the cluster with the highest association frequencies. Each cell's *promiscuity* was subtracted from its stability to calculate cell score.

2.3.8 Implementation

BEARscc exists an R package. Estimation of technical noise from ERCC spike-ins and computation of relevant parameters in order to generate noise injected count matrices is accomplished with the user function,

`estimate_noiseparameters()`.

Noise-injected count matrices are generated using the user function,

`create_noiseinjected_counts()`.

The user functions,

`compute_consensus()`,

`report_cluster_metrics()`,

`report_cell_metrics()`

are available to generate the consensus matrices from a table of cluster labels for each noise perturbation and to generate cluster and cell assessment metrics discussed in this work. Finally, an additional Rscript,

`HPC_generate_noise_matrices.R`,

is provided to allow users to generate noise injected counts matrices in a high-powered compute environment. BEARscc is currently publicly available as an R package with an accompanying example and readme at:

https://bitbucket.org/bsblabludwig/bearscc.

2.4 Optimizing BEARscc to accurately simulate technical replicates

Before proceeding with the BEARscc algorithm (Section 2.3) I wanted to explore the extent to which simulated technical replicates generated by BEARscc resembled real experimental technical noise. I developed several implementations, which were variants of the previous description, in order to empirically compare the simulated noise to actual technical noise. I obtained data from 48 technical replicate libraries from whole tissue brain data diluted into single cell concentrations as described (Section 2.2.1). ERCC spike-ins were systematically added to each dilute sample well as would be standard in *bona fide* single cell samples. I utilized this data as a positive control in order to compare various implementations of BEARscc and their simulated technical noise to the experimentally measured technical noise in these brain samples. It should be noted that during these experiments with the six implementation methods, I normalized all counts data with the function

`estimateSizeFactorsForMatrix()`

from the package, `DEseq2` (Love *et al.* 2014), applied only to the ERCC subset of transcripts in the counts matrix as described by Marioni *et al.* before estimating parameters with BEARscc (Buettner *et al.* 2015). After injecting noise, counts matrices were transformed by the same `DEseq2` size factors with inverse operation into "raw" noise injected counts tables. I discuss the impact of normalization and the procedure for normalization at length in Section 2.6.

In these experiments on implementation, I varied both the method by which BEARscc estimated the *drop-out recovery distribution* and the presence of zero, one, or two functions that appropriately inject dropouts into the data when generating simulated technical replicates using the *drop-out injection distribution* described previous (Section 2.3.4). Thus, in the end I compared six different implementations of BEARscc to determine which most accurately modeled the technical noise observed in the whole tissue brain technical replicates. The finalized algorithm previously described in the methods utilizes a single dropout function and computes dropout noise on a gene by gene basis; I refer to this implementation as the "Genewise - Single dropout function" for the duration of this section. The algorithm description in Section 2.4.1 illustrates the changes in implementation for the "No dropouts" and "Two dropout functions" iterations of BEARscc. The "No dropouts" implementation removes the injection of dropouts entirely (highlighted in red) and the "Two dropout functions" adds an additional, but identical dropout function (highlighted in green) as previously described (Section 2.3.4).

The "Global" implementation operates under the same mathematical motivations as previously described for the "Genewise" implementation, except that the estimation of density of the drop-out recovery distribution using Bayes' law does not occur on a gene by gene basis. Instead, Bayes' law (see eq. 2.8) is used to inform the probability density that the actual transcript count was some positive integer, y, given a zero is observed in the dataset. Specifically,

$$\Pr(Y = y \mid X = 0) = \frac{\Pr(X = 0 \mid Y = y) \cdot \Pr(Y = y)}{\Pr(X = 0)}, \tag{2.8}$$

where X is the distribution of observed counts and Y is the distribution of actual transcript counts across the entire dataset. Trivially, the probability that a zero is observed, $\Pr(X = 0)$, was taken to be the fraction of observed zeros in the counts table. As before, $\Pr(X = 0 \mid Y = y)$ was approximated with an ERCC derived probability distribution, and the prior, $\Pr(Y = y)$ was defined with assumptions and empirical observations to allow estimation of the density distribution for the probability that an observed zero should be some count value, $\Pr(Y = y \mid X = 0)$. In this way, the global algorithm only differs in that the bayesian estimation of the drop-out recovery distribution density; the entire counts table is utilized to estimate $\Pr(X = 0)$ rather than estimating the drop-out recovery distribution on a gene by gene basis.

2.4.1 Pseudocode illustrating the variations in implementation for the BEARscc algorithm

Below is the algorithm as previously described (Section 2.3). However, the modifications exclusive to the "No dropouts", the "Two dropout functions", and "Global" implementations are highlighted. The dropout function removed in "No dropouts" is coloured in red and the additional dropout function in "Two dropout functions" is coloured in green. Finally, the subscript. j, delineating a sampling from from a gene-specific distribution is removed in the "Global" implementation, because dropout effects are inferred across the entire counts table instead of on a gene by gene basis. This is illustrated with yellow.

FOR EACH gene, j

  FOR EACH count, c

    IF $c = 0$

      $n \leftarrow$ SAMPLE one count, y, from $\Pr(Y_{\text{j}} = y \mid X_{\text{j}} = 0)$

      IF $n = 0$

        $c \leftarrow 0$

      ELSE

        $c \leftarrow$ SAMPLE one count from $Z(n)$

dropout $\leftarrow TRUE$ with probability, $\Pr(X = 0 \mid Y = c)$
IF *dropout* $= TRUE$
$c \leftarrow 0$
ELSE
$c \leftarrow$ SAMPLE one count from $Z(c)$
ENDIF
ENDIF
ELSE
IF $c \leq k$
dropout $\leftarrow TRUE$ with probability, $\Pr(X = 0 \mid Y = c)$
IF *dropout* $= TRUE$
$c \leftarrow 0$
ELSE
$c \leftarrow$ SAMPLE one count from $Z(c)$
ENDIF
ELSE
$c \leftarrow$ SAMPLE one count from $Z(c)$
ENDIF
ENDIF
RETURN c
DONE
DONE

2.4.2 Gene-wise implementations of BEARscc are superior to global implementations

I assessed the accuracy of noise estimation of the various BEARscc implementations by comparing the simulated counts from each implementation to the observed counts in the brain technical replicates. Specifically, I examined the resulting mean, variance, and dropout rates for each gene in simulated technical replicates

compared to experimental brain technical replicates in order to determine the most accurate implementation of the algorithm.

Gene-wise implementations better model the the coefficient of variances

In order to understand how the variance normalized by the mean varied across means in the simulated and the technical data for each of the six implementations, I examined the coefficient of variance for each implementation as function of observed mean expression (Figure 2.4.1) for the observed technical replicate data and the simulated data resulting from each BEARscc implementation executed 100 times. I found that all three gene-wise implementations performed better than global implementations in terms of this metric. Specifically, global implementations underestimate the coefficient of variance for low expressing genes and overestimate the variance for high expressing genes. Because global implementations apply the same Bayesian model for recovering dropouts in generating simulated data uniformly across all genes, it is not surprising that the simulated datasets from this set of implementations is incorrectly distributed for genes expressed well above or below the median expression level of the dataset. Amongst the three gene-wise implementations, there was no meaningful difference between the three drop-out function implementations. Similarly, altering the drop-out implementations amongst the three global implementations did not reveal any obvious differences in the distribution of the coefficients of variance as a function of mean expression.

I also observed that the ERCC spike-in transcripts' coefficient of variance relationship with mean expression is representative of endogenous genes in experimental technical replicates of brain tissue (Figure 2.4.1). This observation was critical, because it demonstrated that the underlying assumption of BEARscc, that technical noise may be modeled with ERCC observations, is a reasonable one. Finally, while the gene-wise implementations appeared to model the relationship between the coefficient of variance and mean in bona fide technical replicates well, there were genes in the experimental replicates that are over- or under-dispersed relative to the simulated data. This represents a limitation of BEARscc.

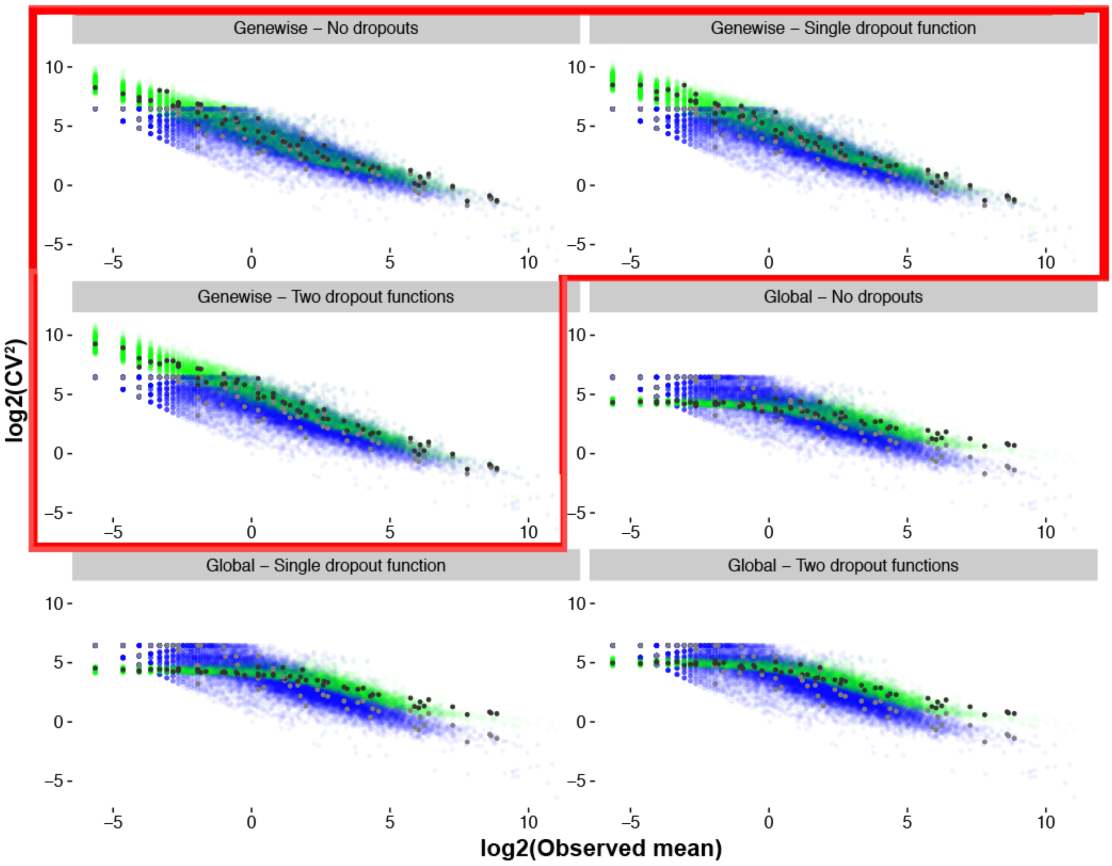


Figure 2.4.1: The genewise implementations of BEARscc (boxed in red) are more accurate in modelling the coefficient of variance as a function of observed mean. Panels represent each of the six aforementioned implementations and colors indicated the gene type: simulated ERCC spike-in (grey), observed ERCC spike-in (black), simulated endogenous (green), or observed endogenous (blue). Panels boxed in red are the more accurate gene-wise implementations.

I also wanted to understand how well the coefficients of variance correlate on a per gene basis between the data simulated by the six implementations of BEARscc and the observed data in the dilute whole tissue brain technical replicates. Therefore, I examined the relationship between coefficients of variation for each gene on a log-scale (Figure 2.4.2). As before, it was clear that the gene-wise implementations were superior to the global implementations in terms of accurately simulating technical noise. While the gene-wise algorithms produced simulated data with coefficients of variance strikingly similar to the actual technical replicates, I found that the global algorithm performed poorly when simulating counts form genes with especially high or low observed coefficients of variation. Furthermore, changing the

number of drop-out functions within the global and gene-wise implementations did not have an obvious effect on the correspondence of genes' coefficients of variance between the simulated and real technical replicates. The strong correspondence between genes' coefficients of variance in simulated and real technical replicates in the gene-wise implementations also emphasizes the degree to which ERCC spike-in measurements are representative of endogenous genes' technical noise operationalized by the coefficient of variance.

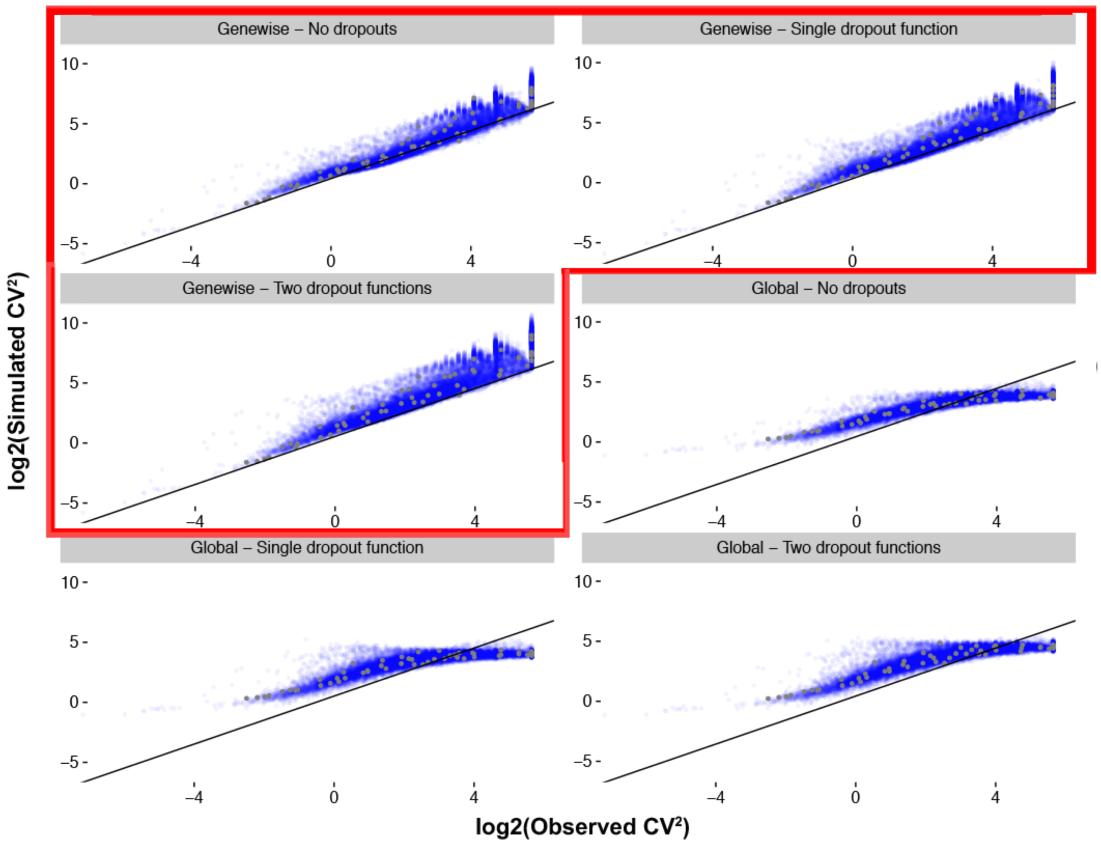


Figure 2.4.2: The gene-wise implementations (boxed in red) of BEARscc more accurately model the coefficient of variance as evidenced by the logarithmic linear correlations. Panels represent each of the six aforementioned implementations and colors indicated the gene type: ERCC spike-in (grey) or endogenous(blue). Panels boxed in red are the more accurate gene-wise implementations.

Gene-wise implementations of BEARscc better model mean and variance

I also examined the accuracy with which each of the six implementations of BEARscc modeled the variance and the mean of the observed technical replicates. First, I compared the mean of each gene across simulated technical replicates to the mean of the same gene in the observed technical replicates (Figure 2.4.3). While the drop-out function choice did not appear to change the fidelity of simulated means to the observed means, the "Global" implementation of BEARscc clearly inflated the mean of simulated replicates compared to observed means for low counts (below 16). In contrast, the "Gene-wise" implementations more closely modeled the mean expression of genes in the observed technical replicates. This is crucial, because a significant distortion of the mean in simulated data would imply that BEARscc biases observations in simulated replicates. Clearly, the "Global" implementations would not be suitable for the final version of BEARscc.

Next, I compared the variance of each gene across simulated technical replicates to the variance of the same gene in the observed technical replicates (Figure 2.4.4). As with mean, I observed that the "Global" implementations were much worse at simulating comparable variance in simulated technical replicates as compared to the observed variance in dilute brain whole tissue samples. Indeed, the "Global" implementations were prone to systematically inflated variance, where the discrepancy only increased in genes with the highest or lowest variance. There were no obvious difference in the correlation of simulated and observed variance between implementations with different numbers of drop-out injection functions. Taken together, these data suggest that "Global" implementations of BEARscc are much worse at simulating technical noise comparable to observed technical noise in experimental replicates.

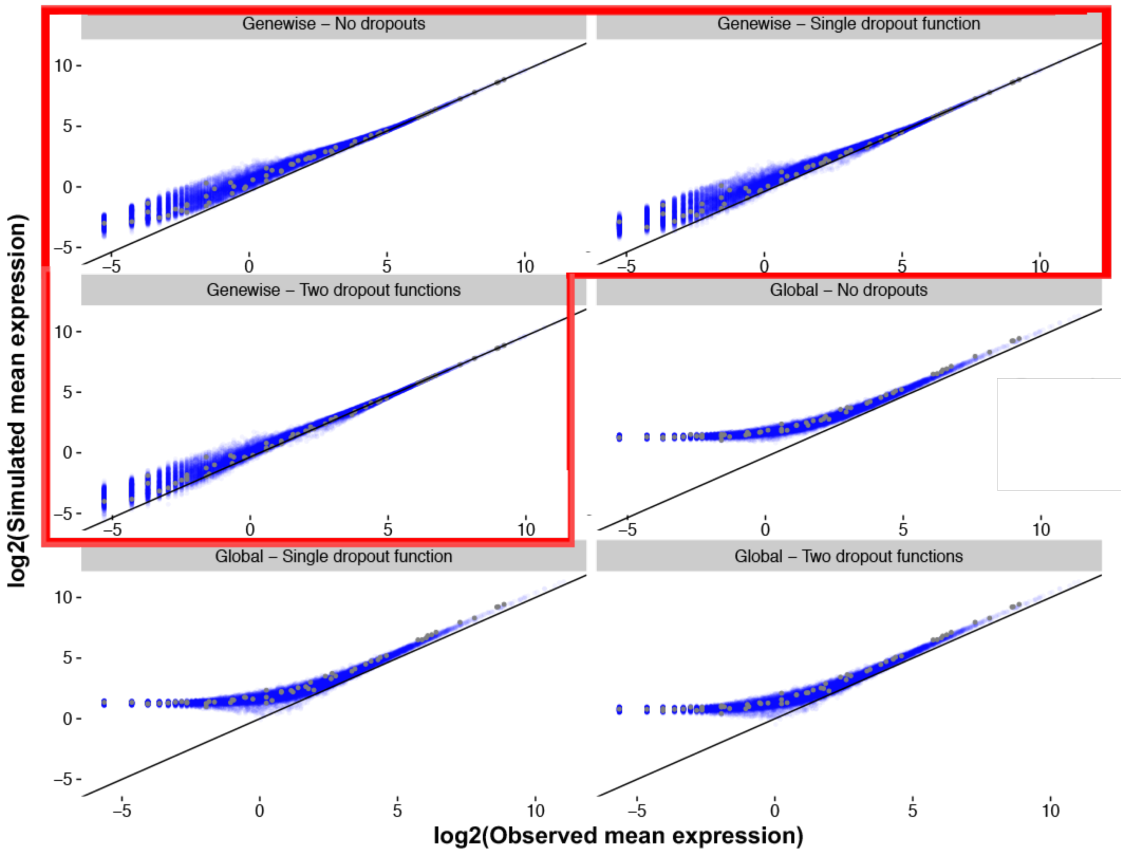


Figure 2.4.3: The genewise implementations of BEARscc more accurately model mean expression as evidenced by the logarithmic linear correlations. Panels represent each of the six aforementioned implementations and colors indicate the gene type: ERCC spike-in (grey) or endogenous (blue). Panels boxed in red are the more accurate gene-wise implementations.

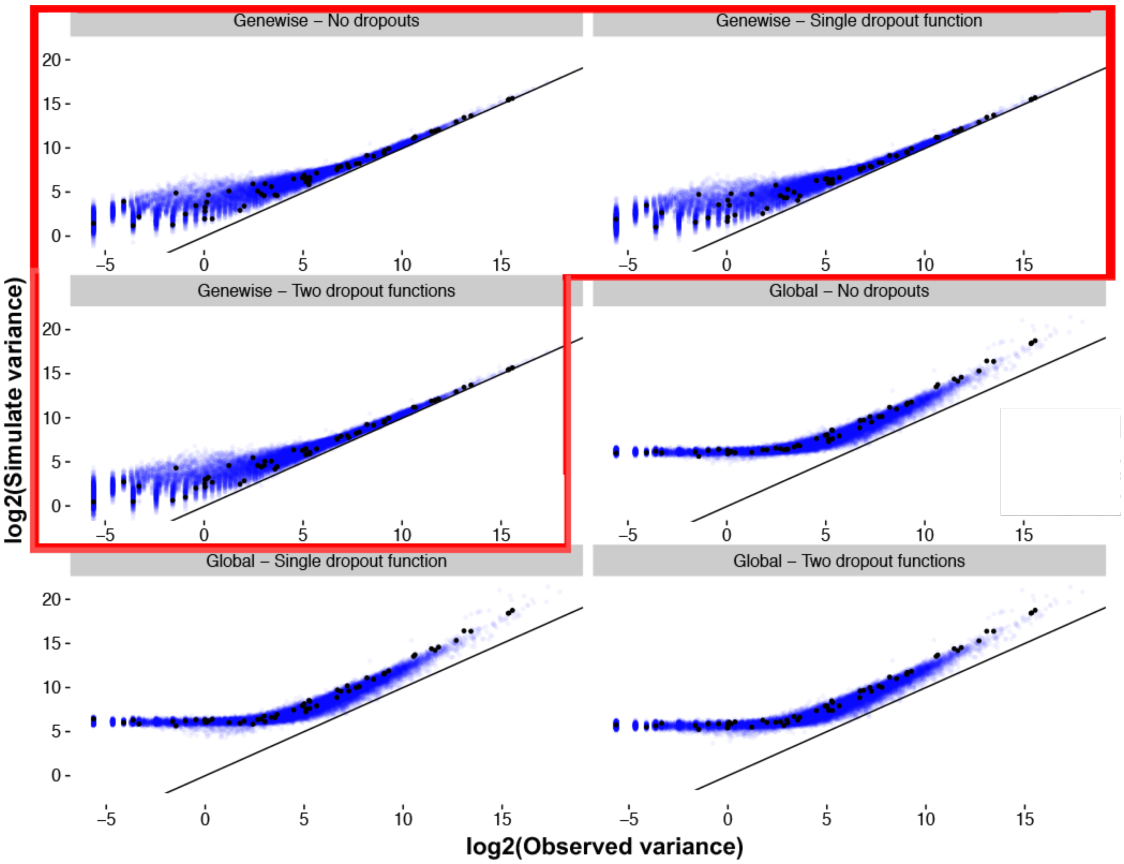


Figure 2.4.4: The gene-wise implementations of BEARscc more accurately model variance as evidenced by the logarithmic linear correlations. Panels represent each of the six aforementioned implementations and colors indicate the gene type: ERCC spike-in (grey) or endogenous (blue). Panels boxed in red are the more accurate gene-wise implementations.

2.4.3 A single drop-out function most accurately models observed drop-out number

In order to understand which implementation most accurately modeled the number of drop-outs observed for each gene, I examined the number of drop-outs on a gene by gene basis in the simulated technical replicates and the observed technical replicates (Figure 2.4.5). As with summary statistics on gene expression, I found that the "Global" implementation generated simulated technical replicates with extensively distorted the the fraction of samples with drop-outs as compared to the observed technical replicates. Specifically, genes with few drop-outs in the observed data had excessive drop-outs in the simulated data, and genes with many drop-outs in the observed data had depleted drop-outs in the simulated data. In contrast, the "Gene-wise" implementations modeled drop-outs reasonably well. The "Gene-wise" implementation with two drop-out functions had systematically inflated drop-outs in simulated replicates for all genes when compared to the observed number of drop-outs. The "Gene-wise" implementation with no drop-out functions generally modeled drop-outs well, but drop-outs in simulated replicates were severely depleted in genes with between 25% and 50%. In contrast, the "Gene-wise" implementation with a single drop-out function modeled drop-outs well in simulated replicates for all genes with observed drop-outs greater than 20%. Additionally, in genes with observed drop-outs below 20%, the simulated replicates' drop-outs per gene were only inflated by approximately 10% on average.

I also wanted to characterize the discrepancy between the simulated and the observed drop-outs for each gene, so I examined the distribution of the genes for each of the six implementations in terms of the discrepancy of drop-out number between simulated and observed drop-outs (Figure 2.4.6). I found that the "Genewise" implementations with no drop-outs and a single drop-out function had the most genes with no change in drop-out number between simulated and observed data. I observed that the single drop-out function had the most balanced distribution between drop-out inflation and depletion in the simulated data, and the vast majority of genes had less than a 25% change in the number of drop-outs between simulated

and observed data for this implementation. Therefore, I concluded that the single drop-out function was the optimal implementation of BEARscc. I also suspected that the discrepancy between drop-outs in simulated and observed data could be improved for BEARscc in future versions of the algorithm.

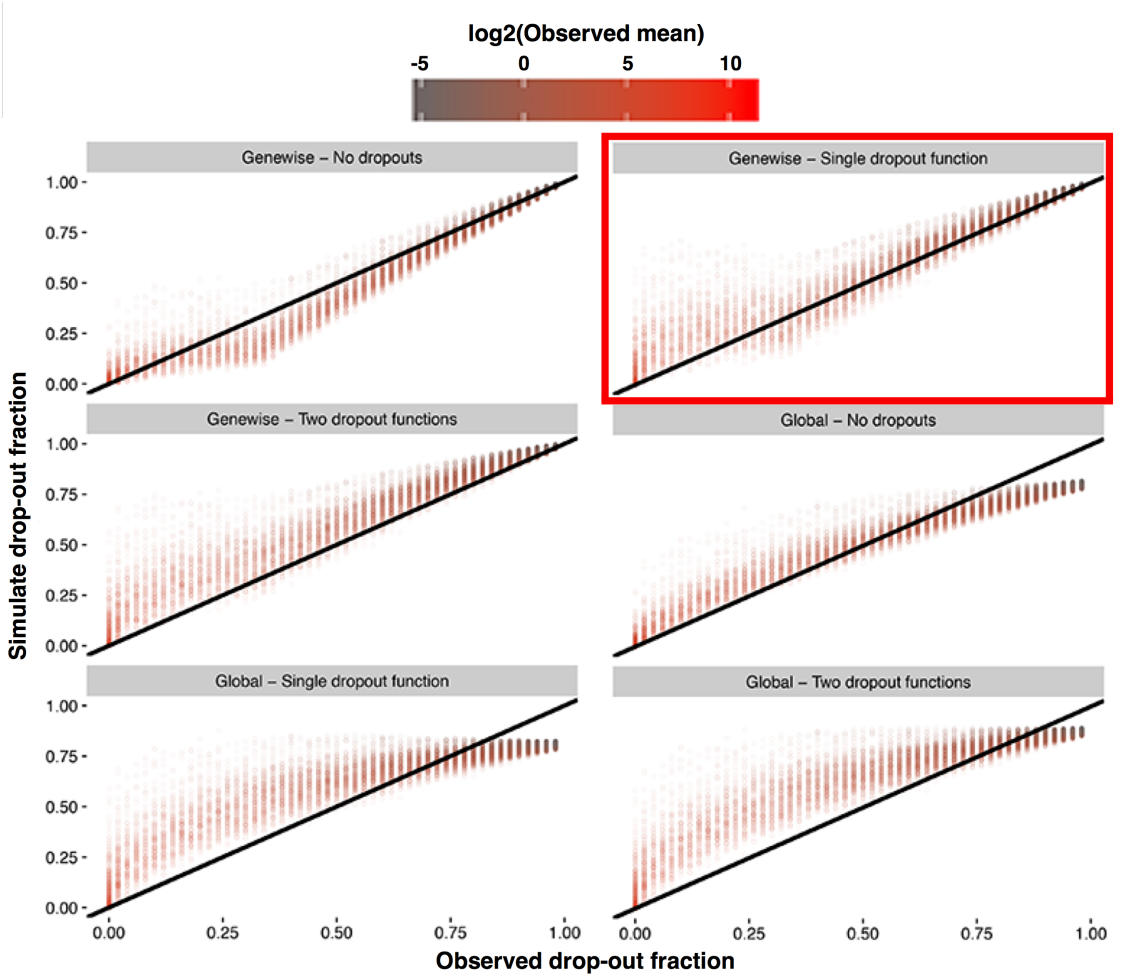


Figure 2.4.5: Scatterplot demonstrates the relationship between the simulated drop-out fraction (y-axis) and observed drop-out fraction (x-axis) , where each point plotted is a gene. The observed expression is indicated by color in log2(mean expression) and the black line denotes a one-to-one relationship. Panel boxed in red is the most accurate drop-out model, the gene-wise single drop-out function.

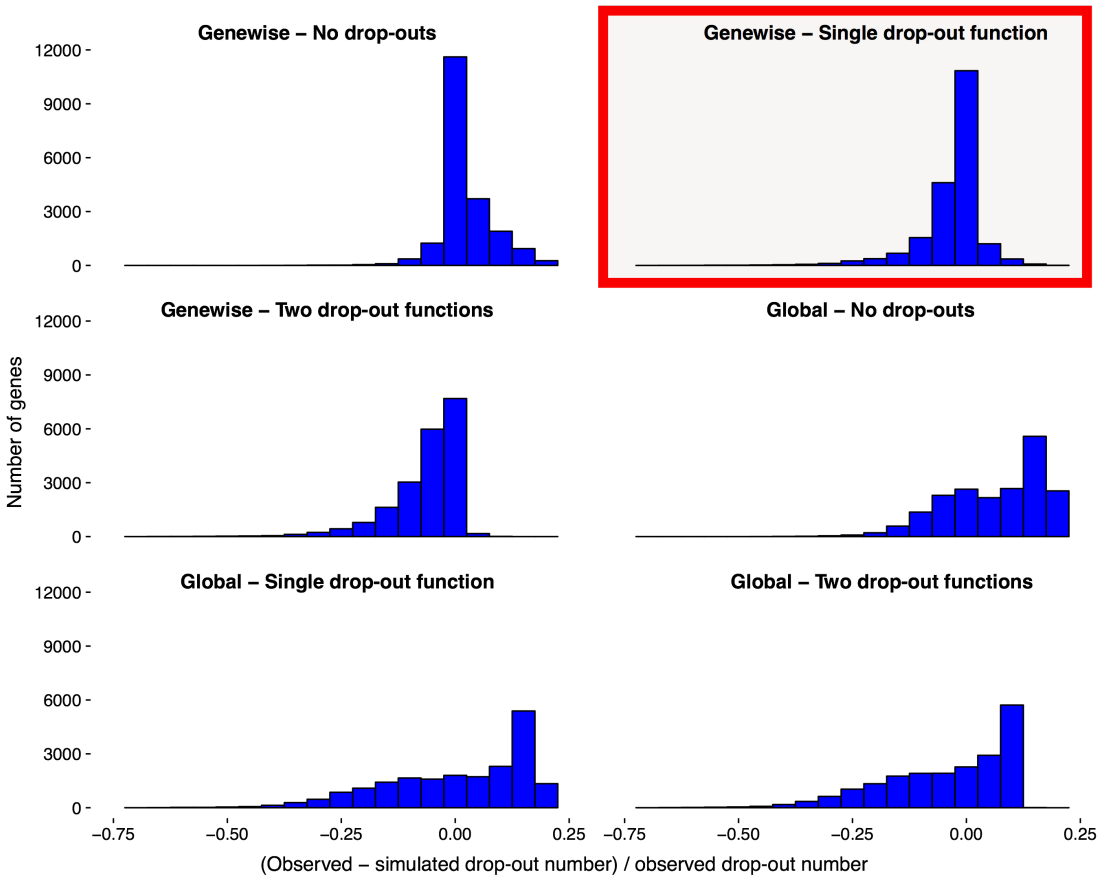


Figure 2.4.6: The genewise implementations of BEARscc are more accurate in modelling the coefficient of variance as a function of the respective mean expression. Panel boxed in red is the most accurate drop-out model, the gene-wise single drop-out function.

In summary, I found that the observed drop-out rate in dilute whole tissue brain technical replicates was most similar to simulated drop-out rates using a single drop-out function. I also observed that gene-wise models of the *drop-out recovery distribution* most accurately modeled several statistical features (coefficient of variance, variance, mean) in experimental technical replicates. Therefore, I established that the most accurate implementation of BEARscc in terms of modelling real technical noise was the gene-wise single drop-out function implementation, which is the implementation described in Section 2.3. All the following analysis in this work was performed with this implementation, and the publicly available R package is this implementation of BEARscc.

2.5 BEARscc accurately models technical noise in single cell RNA-seq data

To assess the accuracy of the noise model employed in BEARscc, we sequenced 45 "blank" samples, which only contained ERCC spike-ins and combined these data with the data from 48 dilute brain samples introduced in Section 2.4. As previously stated, the libraries were otherwise treated identically to a real single cell experiment. This provided an experimental model for measuring the effects of technical noise on single-cell clustering algorithms where two ground truths (brain and blank samples) were clearly defined experimentally.

As noted in Section 2.4, BEARscc simulated technical replicates preserved the mean (Figure 2.5.1a) and variance (Figure 2.5.1b) for transcripts greater than one count when compared to real brain replicates. The simulated mean expression of genes with fewer than an average of 1 observed counts deviated less than 1.5-fold (Figure 2.5.1a). Since expression values below 1 count are outside the dynamic range of any single cell experiment, clusters sensitive to such minute expression differences are unlikely to be reproducible. BEARscc's deviation from expectation in this range should therefore have little practical impact. As previously observed in Section 2.4, BEARscc also preserved the drop-out rate in the simulated data compared to the drop-out rate across genes in the brain technical replicate experiment (Figure 2.5.1c). Specifically, the drop-out rate changed by less than 8.5% for 95% of the genes expressed in the library.

It should be noted that the drop-out rate and the simulated mean value are not perfect, and there is room for improvement. In particular, at low expression levels where the drop-out function is active, we observe inflated simulated mean expression values. It is conceivable that both of these effects are due to the method of inserting the actual observed count values into a random variable model in order to create the simulated counts. Indeed this process assumes that the theoretical mean value of actual transcripts is reasonably close to the observed value, but in low precision experiments this is not true. One mechanism for addressing this would be to examine the observed mean across the experiment and use the distance between

the observed mean and the observed count to weight the observed count to a more feasible value. The downside to this approach is that it assumes that cells will have similar expression of the gene - a patently false assumption in most instance. It is possible that empirical examination could lead to a more clever solution. Barring such an insight, a fundamental change in the paradigm of BEARscc's method may be required to address this issue.

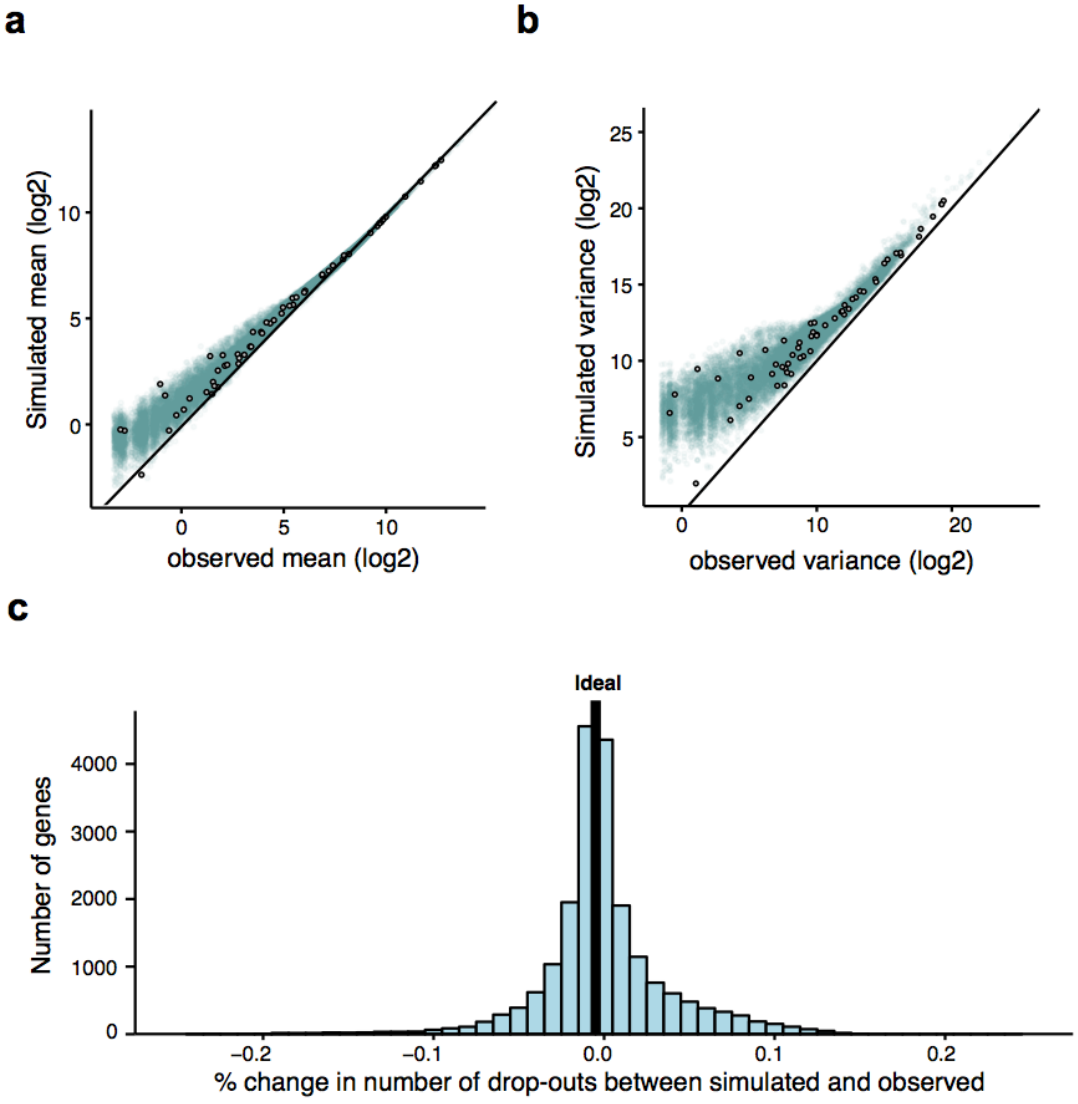


Figure 2.5.1: BEARscc accurately models technical variability. Scatterplots of observed vs simulated mean expression (**a**) and variance in expression (**b**), based on data from brain RNA control experiment. ERCC spike-in values are circled in black, human genes are shown in blue. **c**, Difference between simulated and observed drop-out frequency across genes.

2.6 The importance of normalization in clustering analysis

BEARscc works in conjunction with existing clustering algorithms. To demonstrate this, I clustered diluted whole tissue brain technical replicate data from two batches using three widely used tools (RaceID2 (Grün *et al.* 2016b), BackSPIN (Zeisel *et al.* 2015), and SC3 (Kiselev *et al.* 2016)). In each case, I followed the algorithm-specific recommendations regarding data pre-processing and parameters. However, I was concerned about the role of batch effects on clustering by these algorithms and the resulting analysis by BEARscc. Indeed, work performed by Hicks *et al* illustrated the dramatic impact of batch effects on single cell RNA-sequencing analysis (Hicks *et al.* 2015). Thus I chose to investigate the results of this clustering using four different normalization methods.

I selected four common methods for normalizing data (or lack thereof) with respect to library size and/or sequencing depth: no normalization, DEseq2 normalization, counts per million (cpm) normalization, and adjusted cpm normalization. No normalization was accomplished by simply clustering on the raw counts and cpm normalization was accomplished by dividing the counts per gene for each gene by the total number of counts in the library multiplied by 10^6. DEseq2 normalization was accomplished with the DEseq2 R package (Love *et al.* 2014), whereby a size factor was estimated using only the ERCC spike-in transcripts for each library and all genes within a given library were scaled by this size factor. It should be noted that this normalization is not a fair assessment of DEseq2, and the results should not be taken as such. In fact, the purpose of this was to compare the effects of normalizing only by sequencing depth to those which actually account for total transcription (such as cpm). In hindsight, it would have been ideal to include a proper DEseq2 normalization that depended upon the entire library rather than simply the ERCC spike-ins for this analysis. Finally, the adjusted cpm normalization was accomplished using the approach suggested by Hicks *et al.* I applied three widely used clustering methods (SC3, BackSPIN, and RaceID2) to the whole tissue brain samples from two distinct batches normalized in the four ways described above.

I applied BEARscc to each normalization/clustering algorithm combination in order to assess the algorithm's ability to dissolve batch effects as a function of normalization methods. Upon applying BEARscc in conjunction with RaceID2 (Grün *et al.* 2016b), I found that batch effects were strongly present in both the initial clustering by RaceID2 and the BEARscc noise consensus irrespective of normalization method (Figure 2.6.1a-d). Notably, the extent to which cluster were confounded by batch in the initial RaceID2 clustering remained unchanged for all four normalization methods (Figure 2.6.1a-d, top). However, when BEARscc was applied to RaceID2, there were clear differences in the degree to which the noise consensus displayed association across batches (Figure 2.6.1a-d, top). Specifically, the cpm and adjusted cpm normalization were equally superior in terms of dissolving batch effects as compared to the no normalization and DEseq2 methods. Intriguingly DEseq2 normalization solidified the batch confounded clustering as compared to no normalization. I suspected that this was a direct result of the dramatic difference in library sizes between the two batches (annotation bar in Figure 2.6.1), where the DEseq2 scaling factor calculated from ERCC spike-ins only served to magnify the batch effect. Furthermore, I observed, following further experimentation, that these batch effects could be resolved with adjusted cpm normalization provided that the highest expressing genes with the lowest coefficients of variance were removed from the library (see section 2.7 and Figure 2.7.5).

Upon applying BEARscc in conjunction with BackSPIN (Zeisel *et al.* 2015), I observed that the numerous clusters identified by BackSPIN were often confounded by batch and that the case with no normalization was most strongly confounded by batch (Figure 2.6.2a-d, top). The extent to which the initial clustering by BackSPIN was confounded for batch for the other three normalization methods was comparable. However, only the two cpm based normalization methods were able to dissolve batch effects in conjunction with BEARscc (Figure 2.6.2a-d, bottom). Again, this result is likely an effect of the dramatic difference in library size between the two batch of dilute whole tissue brain data (annotation bar in Figure 2.6.2).

There was no clear difference in the performance of adjusted cpm and traditional cpm normalization in the ability of BEARscc to dissolve batch effects.

Upon applying BEARscc in conjunction with SC3 (Kiselev *et al.* 2016), I found that when no normalization was applied the initial clustering and the BEARscc noise consensus were nearly completely confounded by batch (Figure 2.6.3a). Intriguingly, the DEseq2 normalization resulted in a heavily skewed clustering with only 5 cells in SC3 cluster 2; the BEARscc noise consensus did not perturb the initial clustering at all (Figure 2.6.3b). However, there was a strong association between the 5 cells forming SC3 cluster 2 and a subset of cells in SC3 cluster 1. These clusters were not confounded by batch. I found it intriguing that the DEseq2 normalization failed to dissolve batch confounded clustering when BackSPIN and RaceID2 were utilized as clustering methods, but did appear to dissolve batch effects with the application of BEARscc when SC3 was utilized as the clustering method. The initial clustering by SC3 on the two different cpm normalized dilute whole tissue brain data certainly exhibited bias related to the batch of the cell; however, this bias was not as dramatic as that observed in the initial clustering performed by backSPIN and RaceID2 (Figure 2.6.3c-d, top). When BEARscc was applied to the SC3 clustering of the two types of cpm normalized data, I found that the artifactual clusters were dissolved in both cases (Figure 2.6.3c-d, bottom). However, qualitatively the association frequencies of cells in the adjusted cpm normalization were much higher across the two initial SC3 clusters on average than in the traditional cpm normalization.

Holistically, the adjusted cpm normalization described by Hicks *et al* and regular cpm normalization conjunction with BEARscc were the most successful in dissolving batch effect across the three clustering algorithms. This was especially true for the BackSPIN and the SC3 algorithms, where these where a combination of BEARscc simulation of technical replicates and either form of cpm normalization was required to dissolve batch effects. Furthermore, adjusted cpm normalization qualitatively appeared to more strongly dissolve batch effects in the SC3 clustering experiment when BEARscc was applied and the work described by Hick *et al* was compelling. Therefore, to avoid artifacts due to differences in amplification-dependent library

size, I applied an adjusted cpm normalization to the brain and blank libraries (Hicks *et al.* 2015). Importantly, I suspect for libraries generated with unique molecular identifiers, this step can be omitted because the chemistry in those protocols is less prone to the batch effects observed here. Consequently, I do not apply adjusted cpm normalization to libraries generated with unique molecular identifiers in section 2.9.

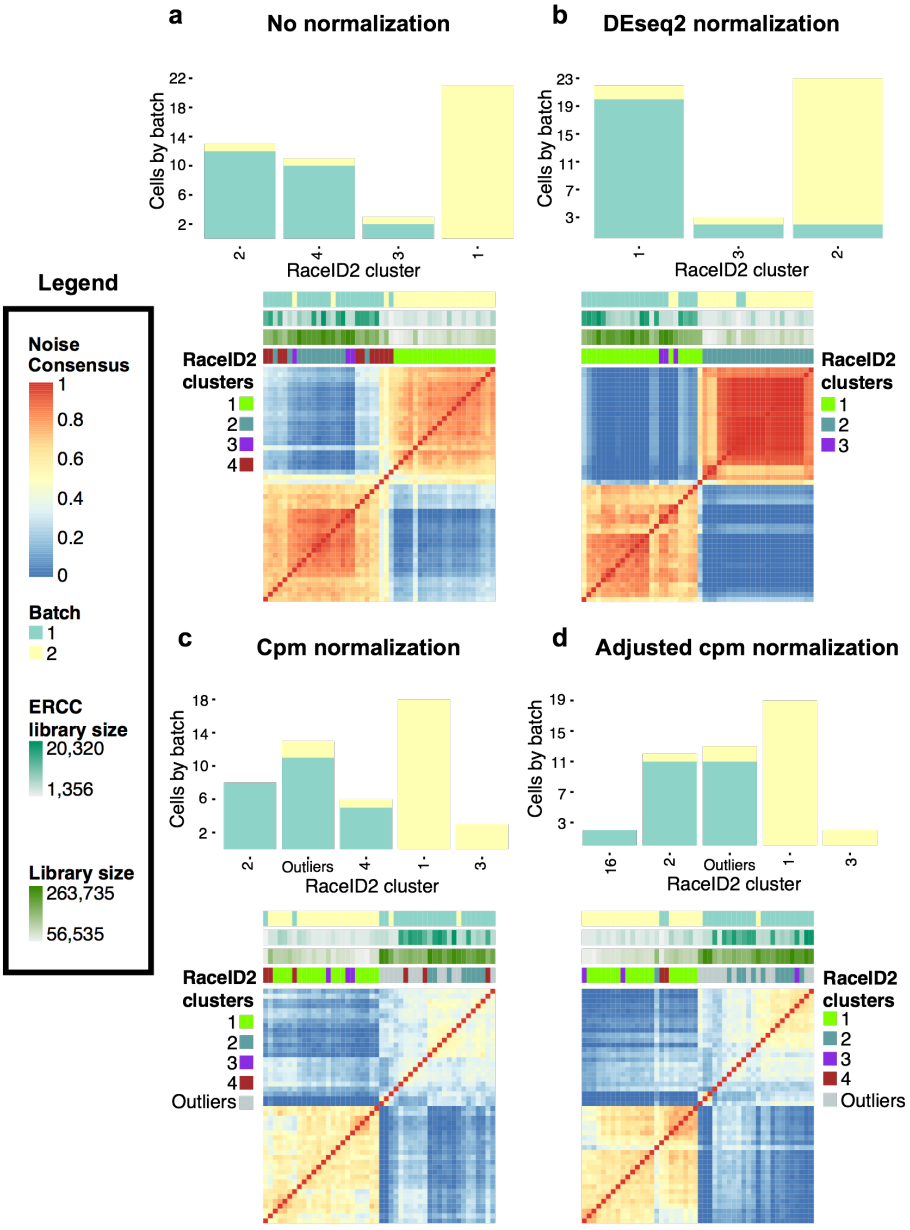


Figure 2.6.1: An adjusted cpm normalization suggested by Hicks et al. most effectively accounts for batch effect noise when using BEARscc applied to RaceID2 clustering (Hicks *et al.* 2015). Bar plots (top) illustrate the degree to which RaceID2 clustering is confounded by batch effects for each normalization method (no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**)), where the number of samples from each batch is indicated by color for each RaceID2 cluster. Heatmaps (bottom) illustrate the noise-consensus matrices of BEARscc applied to the RaceID2 clustering algorithm run on dilute whole tissue brain data prepared in two batches, where each heatmap denotes a different normalization method: no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**).

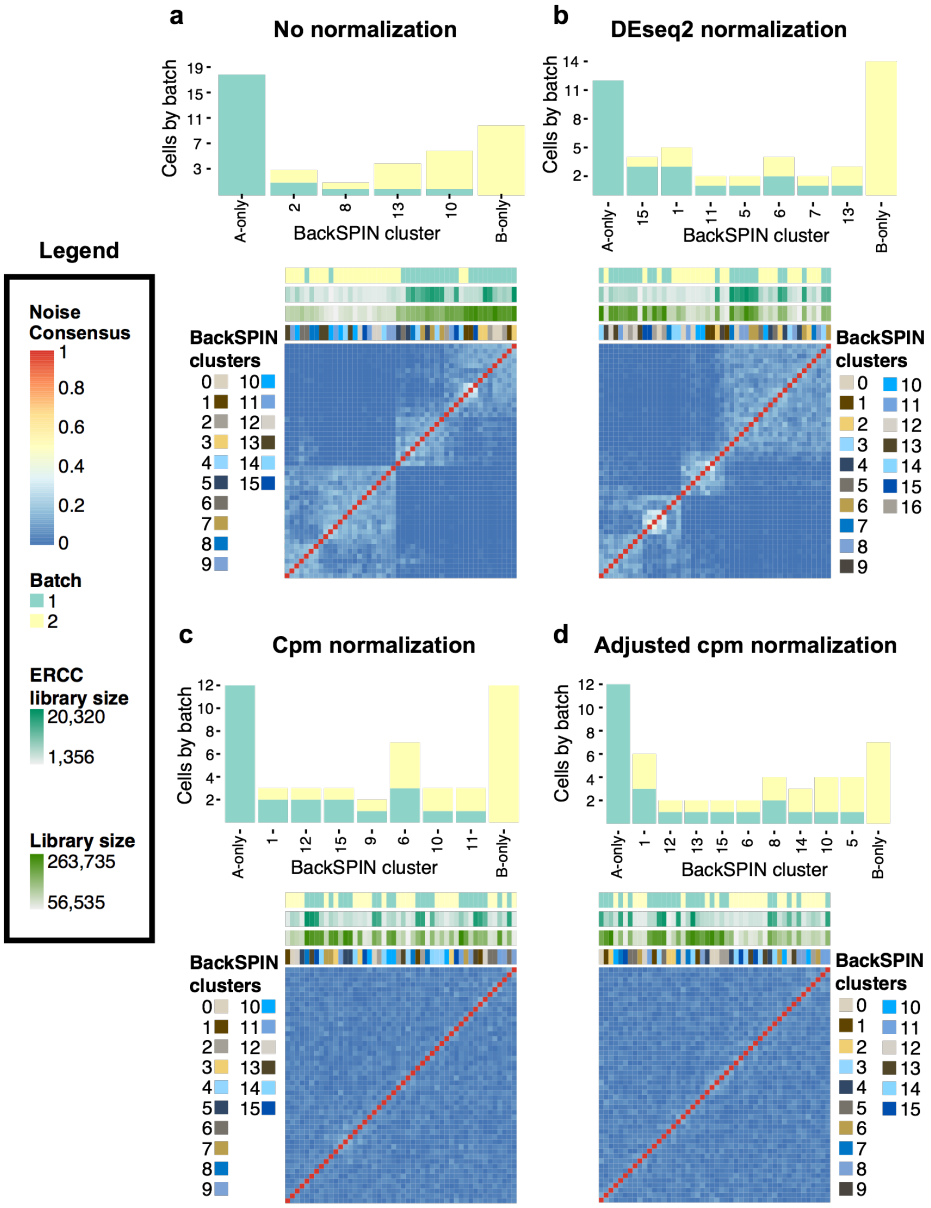


Figure 2.6.2: An adjusted cpm normalization suggested by Hicks *et al.* most effectively accounts for batch effect noise when using BEARscc applied to BackSPIN clustering (Hicks *et al.* 2015). Bar plots (top) illustrate the degree to which BackSPIN clustering is confounded by batch effects for each normalization method (no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**)), where the number of samples from each batch is indicated by color for each BackSPIN cluster. Completely confounded clusters are aggregated together. Heatmaps (bottom) illustrate the noise-consensus matrices of BEARscc applied to the BackSPIN clustering algorithm run on dilute whole tissue brain data prepared in two batches, where each heatmap denotes a different normalization method: no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**).

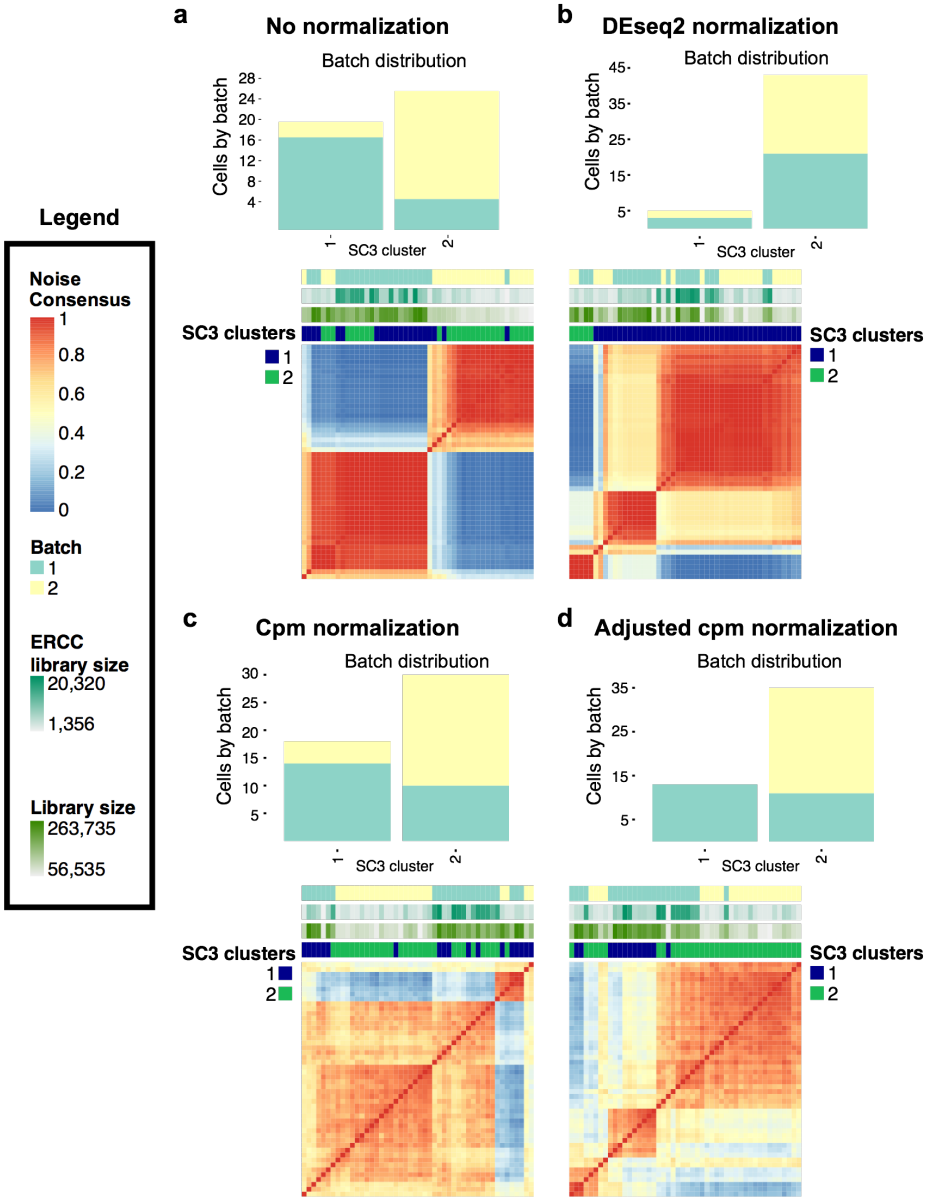


Figure 2.6.3: An adjusted cpm normalization suggested by Hicks *et al.* most effectively accounts for batch effect noise when using BEARscc applied to SC3 clustering (Hicks *et al.* 2015). Bar plots (top) illustrate the degree to which SC3 clustering is confounded by batch effects for each normalization method (no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**)), where the number of samples from each batch is indicated by color for each SC3 cluster. Heatmaps (bottom) illustrate the noise-consensus matrices of BEARscc applied to the SC3 clustering algorithm run on dilute whole tissue brain data prepared in two batches, where each heatmap denotes a different normalization method: no normalization (**a**), DEseq2 normalization (**b**), cpm normalization (**c**), and adjusted cpm normalization (**d**).

2.7 Performance of BEARscc combined with different clustering algrotihms

In addition to assessing the interaction between BEARscc and normalization methods in terms of batch effects, I performed two computational tests using BEARscc in conjunction with the three clustering methods (RaceID2 (Grün *et al.* 2016b), BackSPIN (Zeisel *et al.* 2015), and SC3 (Kiselev *et al.* 2016)) based on our experiment with the diluted brain and blank samples. First, I examined the extent to which BEARscc helped to separate the brain from the blank samples (referred to as the "brain and blanks" experiment). The second test was designed to examine the extent to which BEARscc dissolved batch effects within the brain tissue samples only in adjusted cpm normalized data (referred to as the "batch" experiment). I found that, when used without BEARscc, all three clustering algorithms were prone to create false-positive clusters of samples from the same batch (Figure 2.7.1). Instead of identifying a single cluster of brain samples, I found that SC3 identified 3 clusters on the batch experiment alone (cluster annotation per column in Figure 2.7.1a). Moreover, BackSPIN identified 16 clusters for the batch experiment (cluster annotation per column in Figure 2.7.1e) and RaceID2 identified 3 clusters and 14 outlier clusters in the batch experiment (cluster annotation per column in Figure 2.7.1i). Furthermore, instead of one cluster for blanks and one for brain samples, SC3 identified 4 clusters within the brain and blanks experiment (cluster annotation per column in Figure 2.7.3a) and BackSPIN clustered the data into 28 clusters (cluster annotation per column in Figure 2.7.2e). RaceID2 identified 6 clusters (cluster annotation per column Figure 2.7.2a) instead of the expected two.

For each experiment/clustering algorithm combination, I used BEARscc to generate 100 simulated technical replicates and derived noise consensus clusters as described earlier. For all but one control experiment/clustering method combination, the BEARscc summary statistics highlighted erroneous clusters (Figures 2.7.3, 2.7.1, 2.7.2). I found that for the brain and blanks experiment, BEARscc in conjunction with either SC3 or BackSPIN correctly identified two noise consensus clusters, consisting exclusively of brain tissue or blank samples, respectively (Figure 2.7.3,

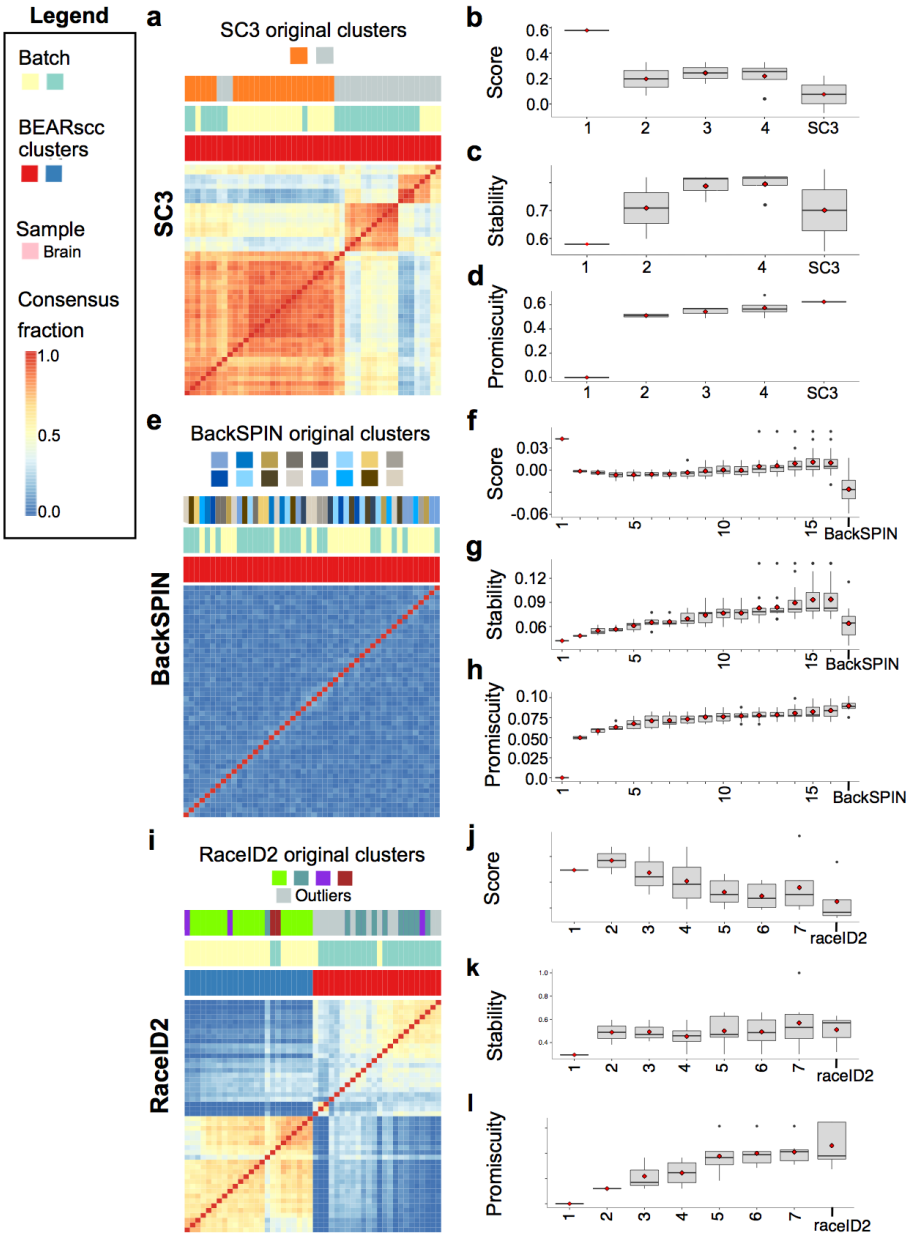


Figure 2.7.1: BEARscc applied to the brain-only control experiment in combination with SC3 (**a**-**d**), BackSPIN (**e**-**h**) and RaceID2 (**i**-**l**). **a**,**e**,**i** Heatmaps illustrate noise consensus clustering resulting from BEARscc applied to SC3 (**a**), BackSPIN (**e**), and RaceID2 (**i**). The putative cluster identity of cells as identified by SC3, BackSPIN, and RaceID2, respectively, are annotated by colour above the heatmap, followed by batch and BEARscc noise consensus clusters. Three boxplots display the BEARscc metrics: *score* (**b**,**f**,**j**), *stability* (**c**,**g**,**k**), and *promiscuity*(**d**,**h**,**l**) for noise consensus clusters with the given set of classes. In each case, the right-most box shows the metric for the original clustering.

Figure 2.7.2f-g). When used in conjunction with SC3 and BackSPIN, BEARscc indicated that one cluster was optimal for the two-batch brain experiment (Figure

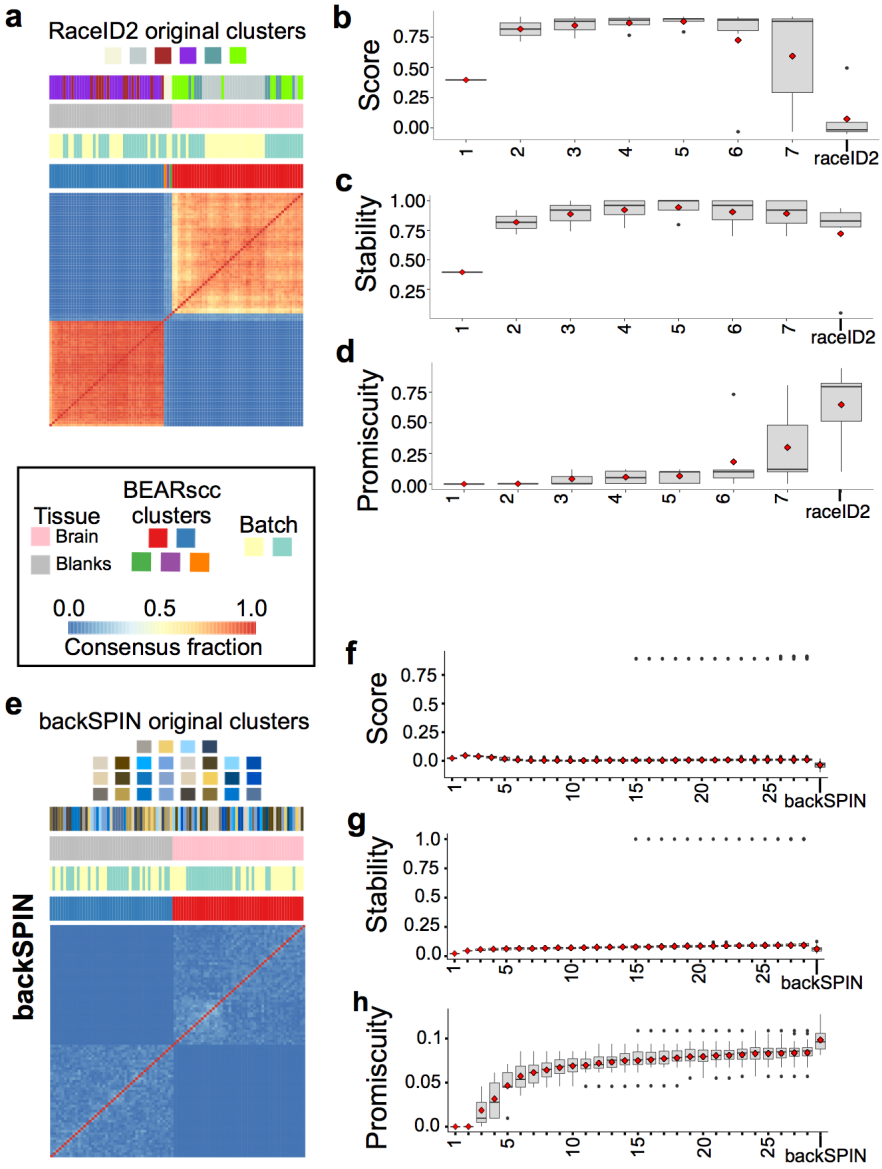


Figure 2.7.2: BEARscc applied to the brain-and-blanks control experiment in combination with RaceID2 (**a**-**d**) and BackSPIN (**e**-**h**). (**a**) Noise consensus matrix resulting from RaceID2 and (**e**) backSPIN. Colour annotations above the heatmaps denote the cluster assignment by RaceID2 and BackSPIN, respectively, tissue, batch, and optimal BEARscc noise consensus clusters. Three boxplots display the BEARscc metrics: *score* (**b**,**g**), *stability* (**c**,**g**), and *promiscuity* (**d**,**h**) per cluster. For RaceID2 clustering the boxplots display the metric for 1 to 7 clusters on BEARscc-RaceID2 consensus matrix and the initial RaceID2 results (**b**-**d**). **e**, Boxplot displays the library size distribution for cells in BEARscc noise consensus clusters containing either one sample (blue) or more than one sample (red). For BackSPIN clustering the boxplots display 1 to 28 clusters on the BEARscc-BackSPIN consensus matrix and the initial BackSPIN results(**f**-**h**).

2.7.1a-h); i.e. the batches did not confound the clustering, which did occur when the clustering algorithms were used alone. In the case of RaceID2 applied to the brain and blanks experiment, BEARscc identified three outlier cells (Figure 2.7.2a-d); the libraries for these three samples contained less than 1,000 observed transcripts (Figure 2.7.4), suggesting that RaceID2's outlier detection feature renders it more sensitive to library size differences than the other tested algorithms.

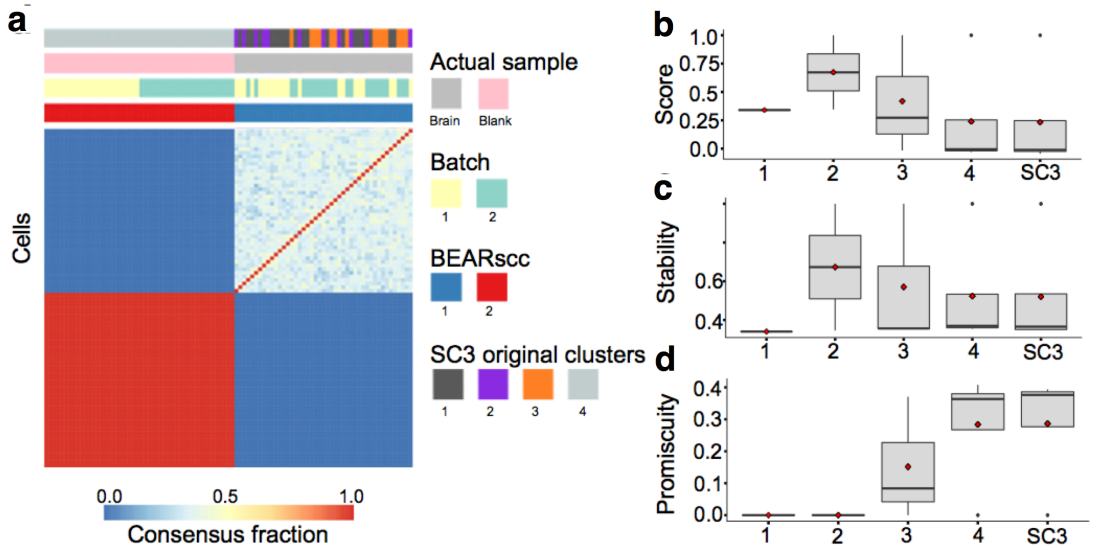


Figure 2.7.3: BEARscc applied to the brain-and-blanks control experiment in combination with SC3 (**a-d**). **a**, Noise consensus matrix of BEARscc applied to SC3 clustering on the brain and blanks experiment. Original SC3 clusters are annotated by colour above the heatmap, followed by sample type, batch and BEARscc consensus clusters. **b-d**, Three boxplots display the BEARscc metrics for evaluating noise perturbations for BEARscc clusterings with 1 to 4 clusters and the original SC3 clustering: *score* (**b**), *stability* (**c**), and *promiscuity* (**d**).

The only case where BEARscc did not fully dissolve a false-positive clustering was when RaceID2 was applied to brain samples only (Figure 2.7.1i-l). In that case, BEARscc identified two clusters as robust to noise; these clusters were almost completely confounded by batch (column annotations in Figure 2.7.1i). Upon further investigation, I found that the batch effect was driven predominantly by very highly expressed genes with low variance: removing the 15 highest expressing genes with coefficients of variance in the bottom fifth percentile resulted in BEARscc identifying only one consensus cluster (Figure 2.7.5a-d). Importantly, RaceID2 alone was unable to resolve batch effects even after filtering (cluster annotation

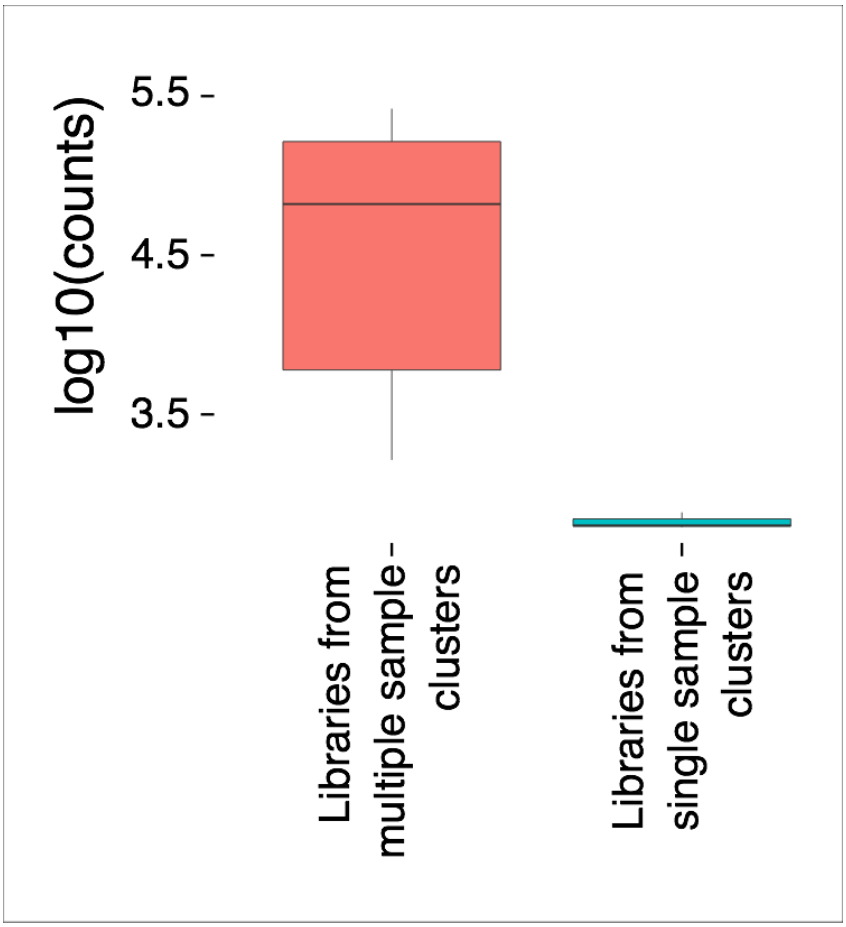


Figure 2.7.4: The three single sample BEARscc noise consensus clusters on the brain and blank libraries are a result of small library size. Boxplots of library size measured in log10(counts) for each sample are displayed, where samples are seperated by libraries in BEARscc noise consensus clusters with more than one sample (red) and only one sample (blue).

per column, Figure 2.7.5a). This suggests that k-medoids clustering on a Pearson correlation matrix, as employed in RaceID2, is sensitive to batch differences in genes with high expression but low variance.

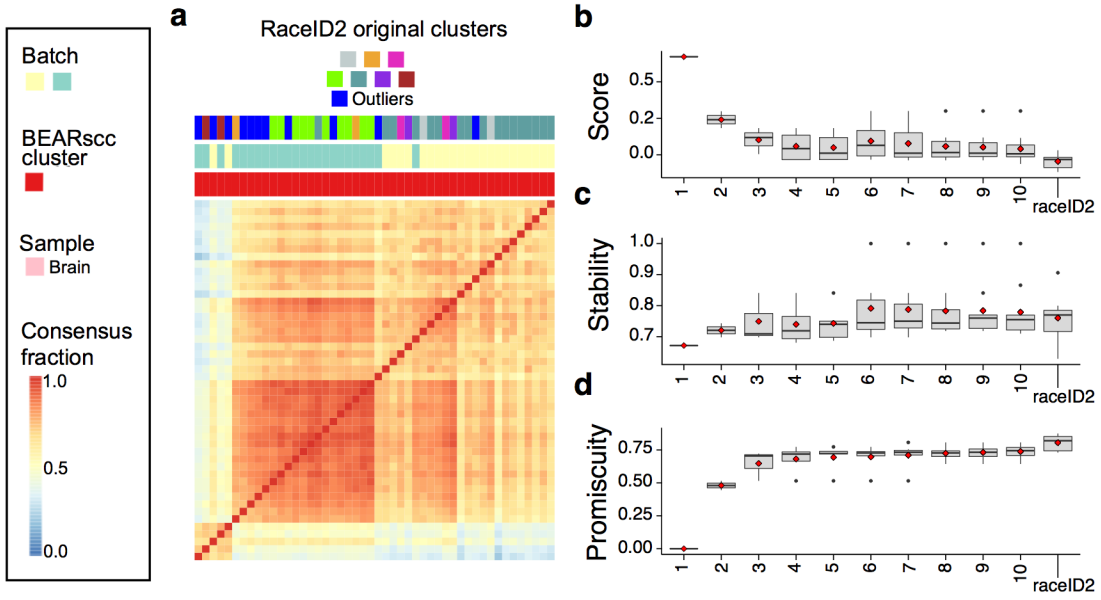


Figure 2.7.5: BEARscc applied to the brain-and-blanks control experiment with filtered genes in combination with RaceID2 (**a-d**). (**a**) Noise consensus matrix resulting from RaceID2 clustering of brain and blank samples, where the 15 highest expressing genes in the bottom 5 percentile of coefficients of variance have been filtered. Colour annotations above the heatmaps denote the cluster assignment by RaceID2, tissue, batch, and optimal BEARscc noise consensus clusters. Three boxplots display the BEARscc metrics: *score* (**b**), *stability* (**c**), and *promiscuity* (**d**) per cluster. The boxplots display the metric for 1 to 10 clusters on BEARscc-RaceID2 consensus matrix and the initial RaceID2 results (**b-d**

2.8 Choosing the number of simulated replicates

I determined the number of simulated technical replicates required to determine the effect of technical variation in the diluted brain data set by testing how often cell-cell associations changed when applying BEARscc to the brain and blanks data, using RaceID2 (Grün *et al.* 2016b) as the clustering tool. I found that the distribution of changes in cell-cell association frequencies in the noise consensus approached zero when the number of simulated technical replicates increases from 50 to 100 and remained clustered at zero when the technical replicate number increases from 100 to 200 (Figure 2.8.1a-b). Therefore, I conservatively elected to use 100 rounds of noise injection for all experiments described below, and recommend that users utilize at least 50 rounds when applying BEARscc.

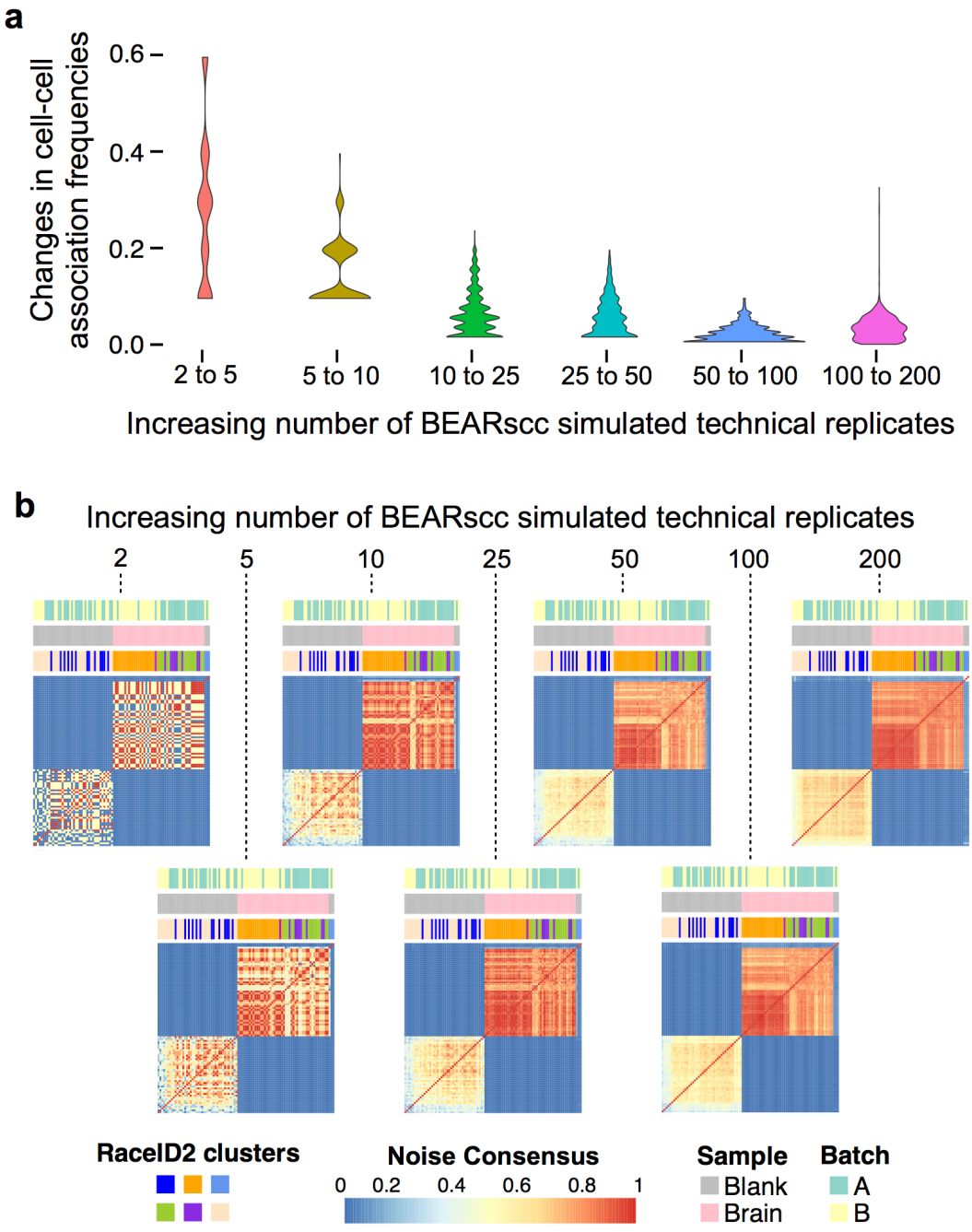


Figure 2.8.1: The effect of the number of perturbations on the noise consensus matrix becomes negligible after 50 perturbations. BEARscc was applied to RaceID2 clustering of brain whole tissue data and blank samples with increasing numbers of simulated technical replicates. **a**, A violin plot illustrating the distribution of cell-cell association frequency changes in the noise consensus matrix as the number of simulated technical replicates increases from 2 to 5 up to 200. **b**, Heatmaps of the noise consensus matrix calculated by BEARscc for each number of simulated technical replicates. Above heatmaps color bars indicate batch (top), sample source (middle), and naïve clustering by RaceID2 (bottom).

2.9 BEARscc identifies cells and clusters that are robust to noise

To assess the utility of BEARscc in a real-world scenario, I applied the algorithm to two publicly available datasets: murine intestinal data from a lineage tracing experiment obtained by Grün *et al.* (Grün *et al.* 2016b) and murine brain data obtained by Zeisel *et al.* (Zeisel *et al.* 2015).

2.9.1 An illustration in murine intestine single cell samples

When I ran the RaceID2 algorithm developed by Grün *et al.* on the un-altered mouse intestinal cell data, the cluster annotations for 45 out of 291 cells differed from the published results, purely due to the stochastic nature of the underlying k-medoids algorithm (Figure 2.9.1a-b). To distinguish this algorithmic stochastic variation from the effect of simulated technical variability generated by BEARscc, I ran RaceID2 100 times on the observed, un-altered expression counts and compared the resulting consensus matrix (Figure 2.9.1c) to the noise consensus matrix derived from re-clustering 100 simulated technical replicates produced by BEARscc (Figure 2.9.2a). The BEARscc noise consensus matrix differs substantially from the consensus matrix of the algorithmic variability (Figure 2.9.2a and 2.9.1c, respectively), with BEARscc resulting in more discrepancies in cell-cell associations than the k-medoids consensus (Figure 2.9.1d). This confirms that simulated technical variation as generated by BEARscc has a substantially larger and qualitatively different impact on the resulting clustering than the non-deterministic nature of the RaceID2 algorithm alone.

Applying RaceID2 to the mouse intestinal data, Grün *et al.* described 5 main clusters (referred to as R1-R5) and 17 rare cell type clusters (R6-R22), of which 7 consisted only of a single cell (Grün *et al.* 2016b). Grün *et al.* originally developed RaceID2 to identify rare cell types and to identify developmental lineage trees in differentiating tissue. The algorithm therefore tries to strike a balance between detecting more rare clusters while trying to avoid over-clustering. However, RaceID2 provides no metric to evaluate how technical variability between experiments could affect the clustering.

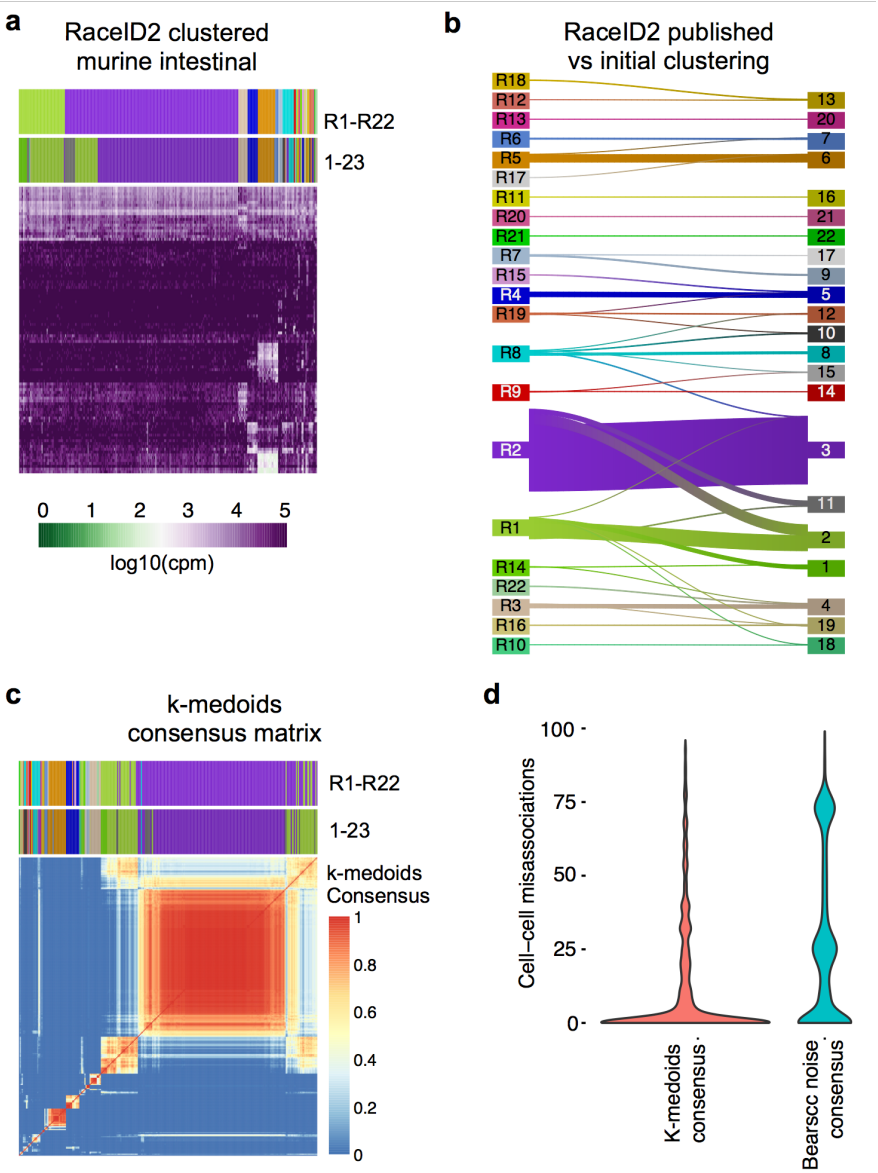


Figure 2.9.1: Clustering of observed counts in mouse intestinal and brain tissue with RaceID2. Published data and re-calculated results do not fully agree. **a**, RaceID2 clustering of read counts of top the 10 most characteristic genes for each cluster as described by Zeisel *et al.* Published clusters and clustering results from our application of RaceID2 are annotated by colour above the heatmap. **b**, River plot displays relationship between published intestinal clusters and our attempt to replicate them. **c**, Heatmap displaying the k-medoids consensus matrix across 100 k-medoid clustering iterations with RaceID2 on the published murine intestinal data. Published cluster results and clustering results from our initial application of RaceID2 are annotated by colour above the heatmap. **d**, Violin plot illustrating the distribution of the number of cell-cell mis-associations relative to the initial clustering of observed counts for the k-medoid consensus and the BEARscc noise consensus.

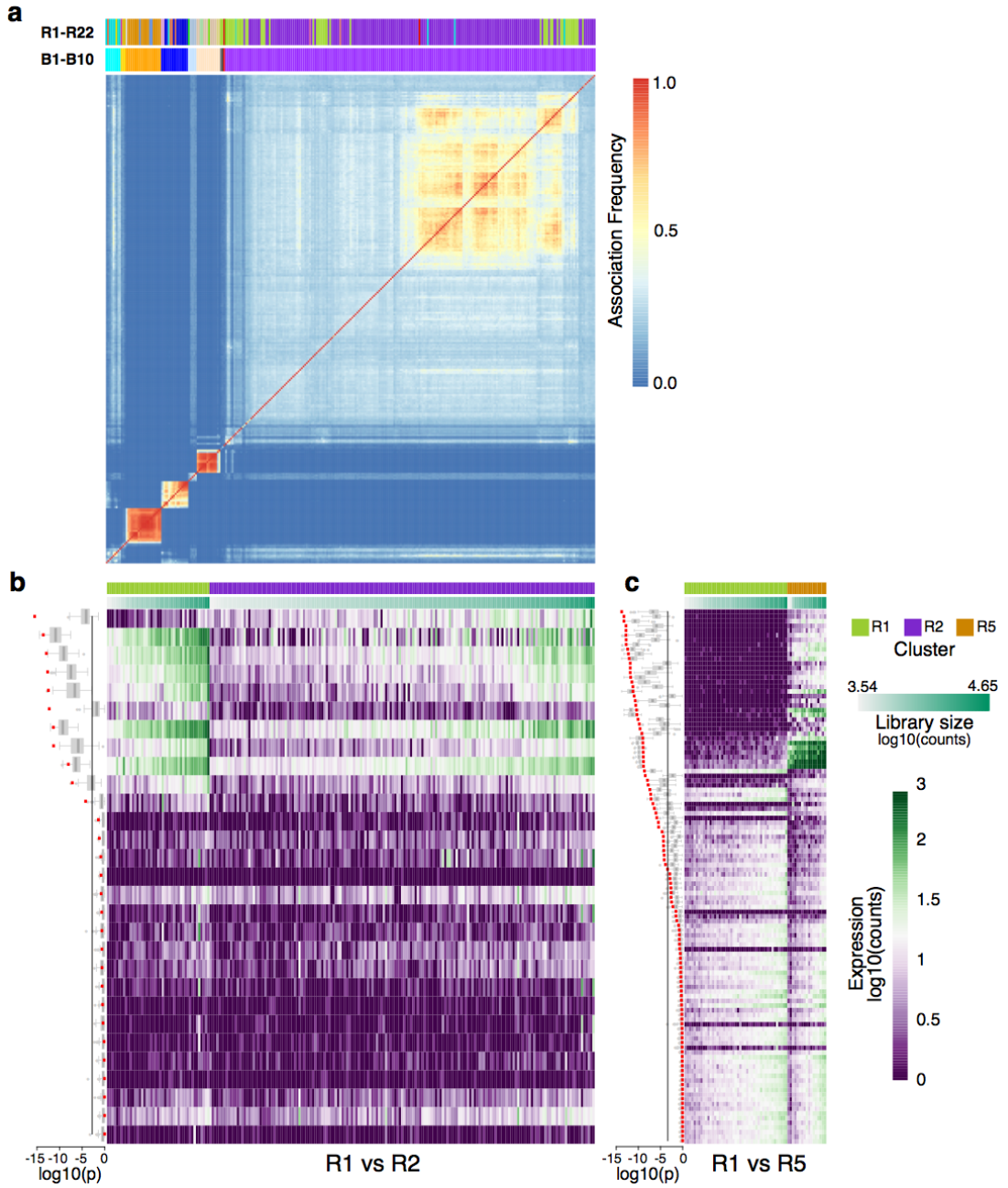


Figure 2.9.2: BEARscc identifies robust cells and clusters in murine intestinal data clustered by RaceID2. **a**, Heatmap displaying the noise consensus resulting from 100 noise-injected counts derived from murine intestinal single cell data published by Grün *et al.* The optimal BEARscc clusters (B1-B10) and the clusters published by Grün et al. (R1-R22) are annotated in colour above the heatmap. **b-c**, Heatmaps of the expression of genes characterizing clusters R1 and R2 (**b**), and clusters R1 and R5, as described by Grün *et al.* (**c**). Columns ordered by library size per cell, rows sorted by significance of expression fold-change between clusters. The p-values (Wilcoxon rank-sum test) for each gene are indicated in red (observed values) and in grey (simulated technical replicates) to the left of the heatmaps. Solid line denotes Bonferroni-corrected significance threshold.

After simulating technical replicates with BEARscc and re-clustering with RaceID2, 219 out of 291 cells were clustered identically in the noise consensus and the published clustering (Figures 2.9.2a and 2.9.1b), suggesting that a large majority of cluster assignments are robust to technical variability. Interestingly, the previously suggested BEARscc heuristic for cluster number (see Section 2.2.3) indicated that 10 clusters were optimal in these data (Figure 2.9.3a). This suggests that some clusters published by Grün *et al.* may be an artifact of experimental noise. Of the 5 published non-outlier clusters, R3 (enterocytes), R4 (goblet cells) and R5 (Paneth cells) were particularly robust to noise (Figure 2.9.3b). However, groups of cells from clusters R1 to R5 frequently associated with cells in several outlier clusters across simulated technical replicates. This suggests that RaceID2's selection of outlier cells is frequently based on genes whose expression difference between cells is within the range of technical variability. Simulating technical replicates with BEARscc is therefore important when trying to identify biologically meaningful rare cell populations.

The "stem-like" cluster R2 and the transit-amplifying cluster R1 exhibited markedly lower scores than the other main clusters (Figure 2.9.3b) and present as a single group in the noise consensus matrix (Figure 2.9.2a). Grün *et al.* suggest that R1 and R2 are in fact closely related cell types. To understand the reasons for the low robustness of clusters R1 and R2, I compared the expression of all genes that Grün et al. listed as characteristic for these two clusters (heatmaps in Figure 2.9.2b,c). Expression patterns of these genes were highly similar, compared to those observed between R1 and the next-largest cluster R5. To confirm this, I performed a Wilcoxon signed rank test on each gene, comparing the expression distribution of library size matched cells in clusters R1 and R2 (vertical boxplot in Figure 2.9.2b), and R1 and R5 (vertical boxplot in Figure 2.9.2c). In the case of R1 and R2, the expression fold-change was reduced in technical replicates, relative to that observed in the original data (shown as red bars in Figure 2.9.2b), frequently falling below the significance threshold. These observations suggest that many cells

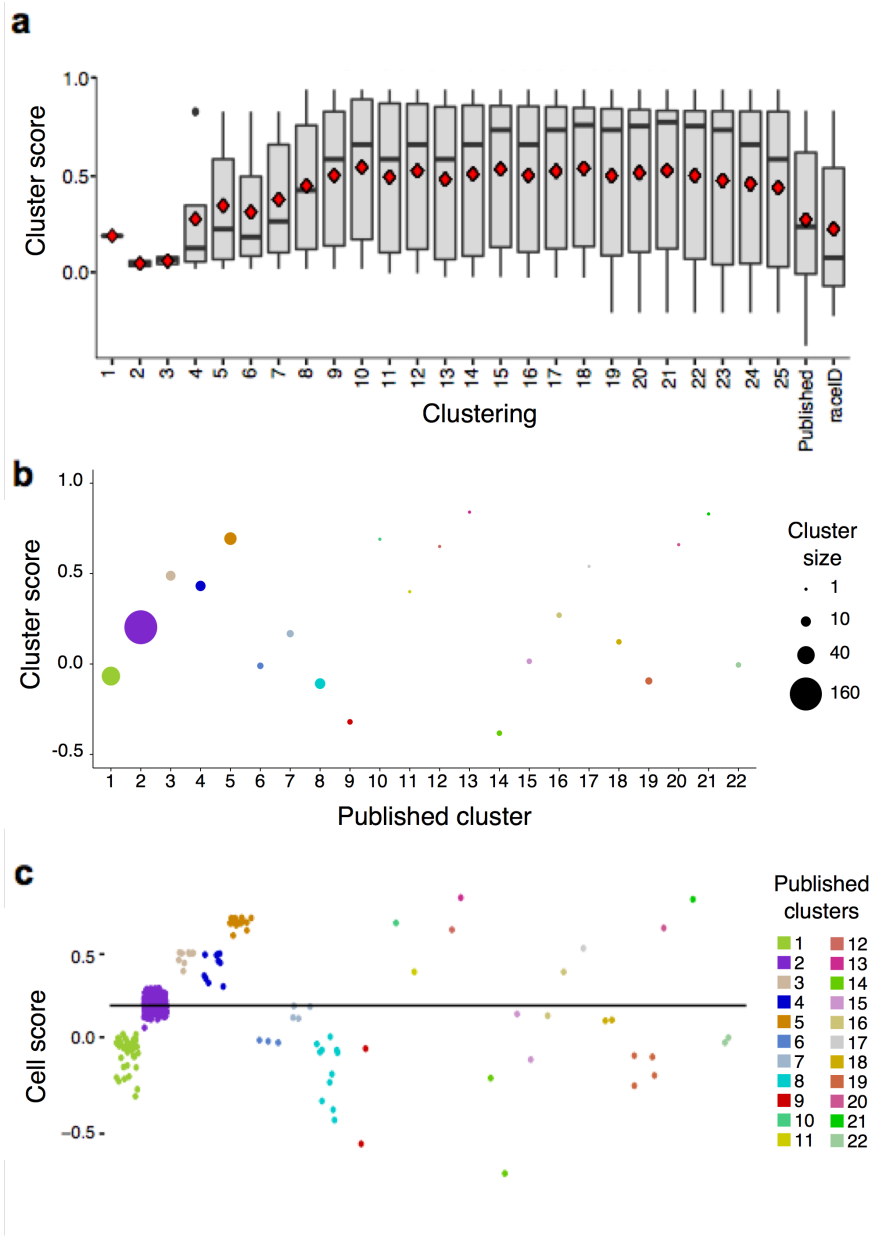


Figure 2.9.3: BEARscc identifies robust cells and clusters in murine intestinal data clustered by RaceID2. **a**, Boxplot displaying the distribution of cluster scores for each potential BEARscc noise consensus clustering with various cluster numbers, k, and for the published RaceID2 clusters and the initial RaceID2 clustering (right). Red diamonds indicate mean values for each clustering. The yellow star indicates the cluster number k with the highest mean cluster score. **b**, Cluster scores for "main" clusters (1-5) and outlier clusters (6-22). Circle size reflects number of cells per cluster. Colors are the same as in subfigure c. **c**, Jitterplot displays the per-cell scores for murine intestinal cells relative to the published RaceID2 clustering. Cells are colored by the cluster to which they were assigned in the origin publication.

in R1 and R2 cannot be reliably classified, but rather appear to represent a point on a gradient of differentiation between two cellular states.

2.9.2 An illustration in murine brain single cell samples

The murine brain data from Zeisel *et al.* were analyzed using the BackSPIN algorithm. In the original publication, the authors post-processed the clustering results to manually group cells into 9 major classes. Unfortunately, only the sequencing data from cells that passed certain filters were made publicly available (Zeisel *et al.* 2015). As a result, I was unable to replicate the published clusters identically. I limited BackSPIN to 5 splits as described by Zeisel *et al.*, which resulted in 24 clusters (Figure 2.9.5a). These were then manually collapsed into 9 main classes based on the same biological interpretation as in the original paper (Zeisel *et al.* 2015) (Figure 2.9.5a). In this way, clustering of 2816 out of 3005 cells was consistent with published results (Figure 2.9.5a,c). Noise consensus clustering using BEARscc dissolved the 24 un-curated clusters (Figure 2.9.5a-b) into 11 reference clusters (B1-B11) (Figure 2.9.5b and 2.9.4). These clusters correspond well with the manually curated clusters. BEARscc's analysis indicates that for each of the broad clusters, a small number of cells appear to have been mis-annotated in the manual curation (Figure 2.9.5b-c). Together these results demonstrate that BEARscc reduces the variability of BackSPIN and highlights real biological cell groups.

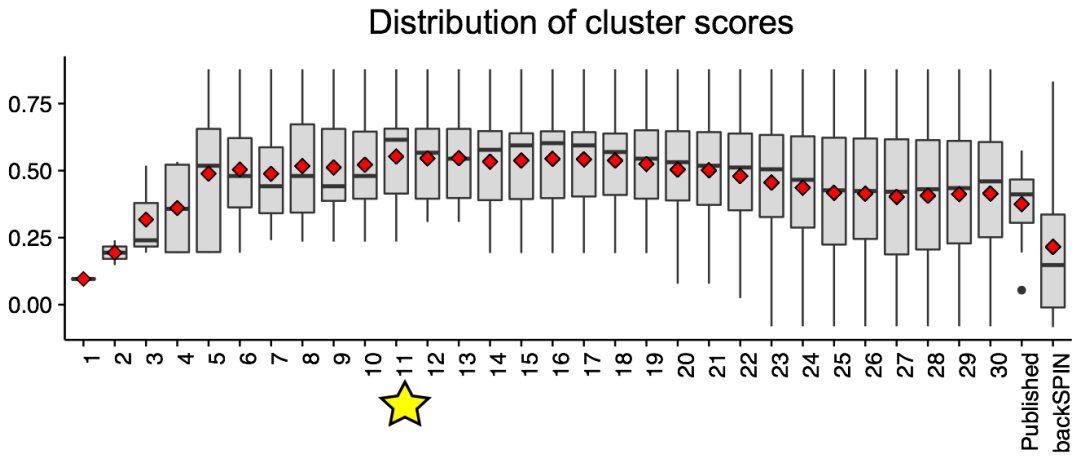


Figure 2.9.4: Boxplot displaying the distribution of cluster scores for each potential BEARscc noise consensus clustering with various cluster numbers, k, and for the published BackSPIN clusters and the initial BackSPIN clustering (right). Mean values for each clustering are indicated by red diamonds, and the chosen k representing the maximum score is denoted by a star.

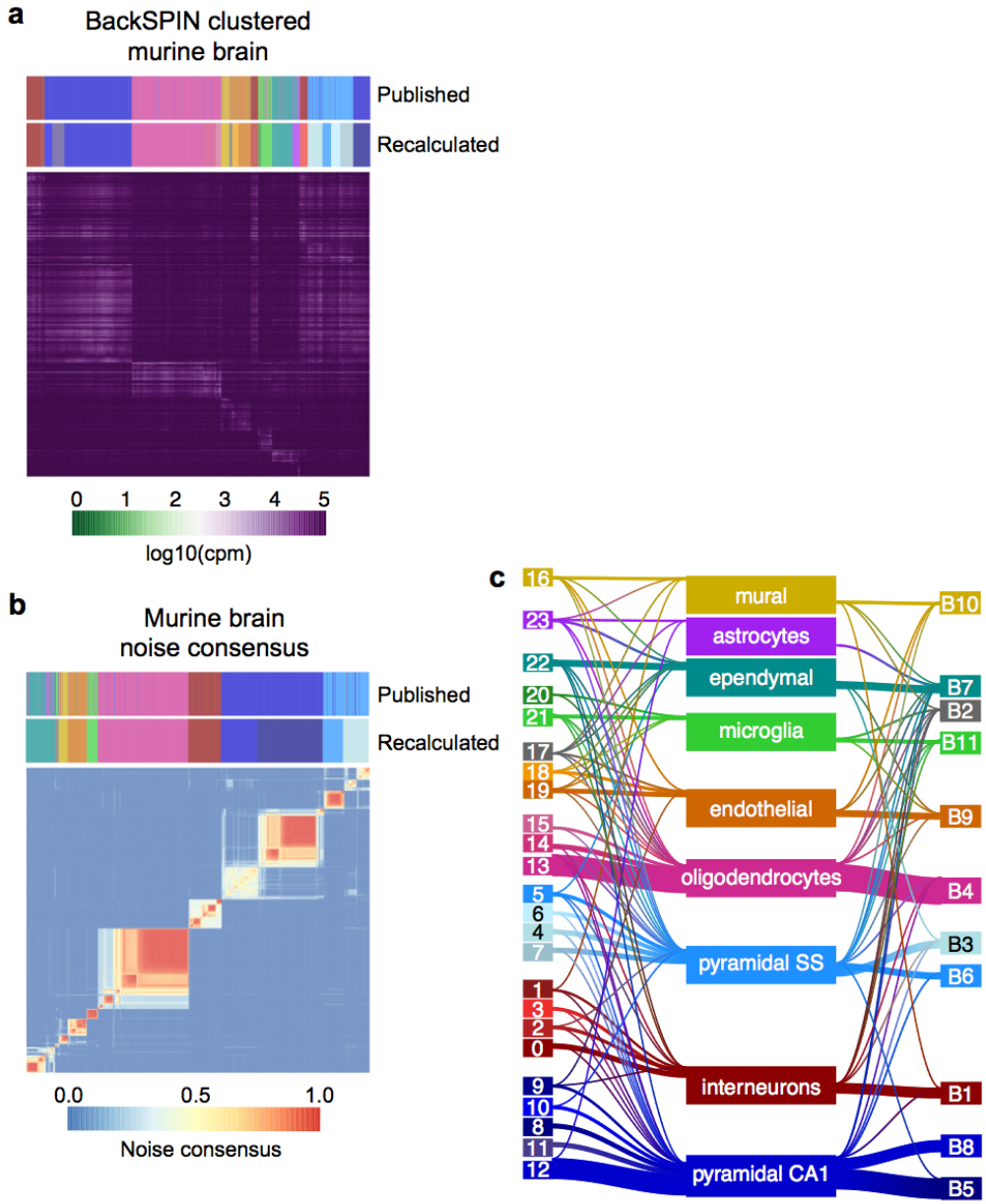


Figure 2.9.5: BEARscc identifies robust cells and clusters in murine brain data clustered by BackSPIN. **a**, BackSPIN clustering on the 5000 genes identified by biclustering from mouse brain data. Published clusters and clustering results from our application of BackSPIN are annotated by colour above the heatmap. **b**, Noise consensus matrix derived from BEARscc applied to BackSPIN clustering of murine brain single cell data published by Zeisel *et al.* The optimal BEARscc clusters and the clusters published by Zeisel *et al.* are annotated by colour above the heatmap. **c**,. River plot showing nodes for each original BackSPIN cluster (left), published BackSPIN cluster (middle) and BEARscc consensus cluster set (right), where the thickness of the connecting line corresponds to the number of cells that are found in the respectively connected original BackSPIN, published, and BEARscc nodes.

2.10 Benchmarking BEARscc against simple gene sampling

In order to benchmark BEARscc, I compared its performance to a simple gene sampling approach (Section 2.2.4). One hundred subsampling iteration matrices for n cells and m genes were generated by sampling one half of expressed genes and clustered using the respective clustering algorithm (SC3, BackSPIN, RaceID2). For each dataset, genes were excluded with less than 25 total raw counts across all samples in the cohort. The remaining genes formed the sample space. In each subsampling iteration, one half of the genes were sampled without replacement from the counts matrix. Identically to the computation of the BEARscc noise consensus matrix, cluster labels were used to compute an n x n binary association matrix for each clustering, and an arithmetic mean was taken for each respective element across the resulting 100 association matrices to produce an n x n subsampling consensus matrix. Identically to BEARscc analysis, the BEARscc score metric was used to determine cluster number k (Supplementary Figures A.1.1, A.1.2, A.1.3). The resulting cluster labels for each dataset and algorithm were compared with BEARscc by computing the adjusted rand index (Rand 1971) for each with respect to the relevant ground truth.

BEARscc outperformed subsampling and naive clustering with the respective each algorithm in the control dataset; critically, this was the only dataset where a definitive ground truth was known (Figure 2.10.1; top). Importantly, when BEARscc was benchmarked on the murine intestine and murine brain using the clustering algorithm that defined the ground truth (RaceID2 and BackSPIN, respectively), BEARscc performed better than gene sampling (Figure 2.10.1; bottom middle and middle left, respectively). In the murine intestine BEARscc performed comparably to RaceID2, which defined the ground truth. In the murine brain BEARscc more closely approximated the curated 9 clusters published with the dataset than the BackSPIN algorithm itself. Notably, in four of the clustering/dataset combinations, BEARscc performed worse than either the algorithm itself, gene sampling, or both. However, because the ground truth for each of these experiments was defined by

either RaceID2 or BackSPIN in the murine intestine and murine brain, respectively, the analysis also reflects a comparison between the algorithms themselves; this confounds the interpretation of the benchmarking considerably. Additionally, by using the cluster number in the ground truth for SC3, I may not have chosen the optimal cluster number parameter k, given SC3 clustering behavior. Most importantly, while benchmarking against gene sampling in this way is a useful mechanism for checking the sensibility of the cluster labels generated by BEARscc, the tool was designed to be used in a more interactive fashion. That is, users should not strictly use BEARscc as another clustering algorithm, but a tool to explore the role of noise in the clustering of their data with their algorithm of choice.

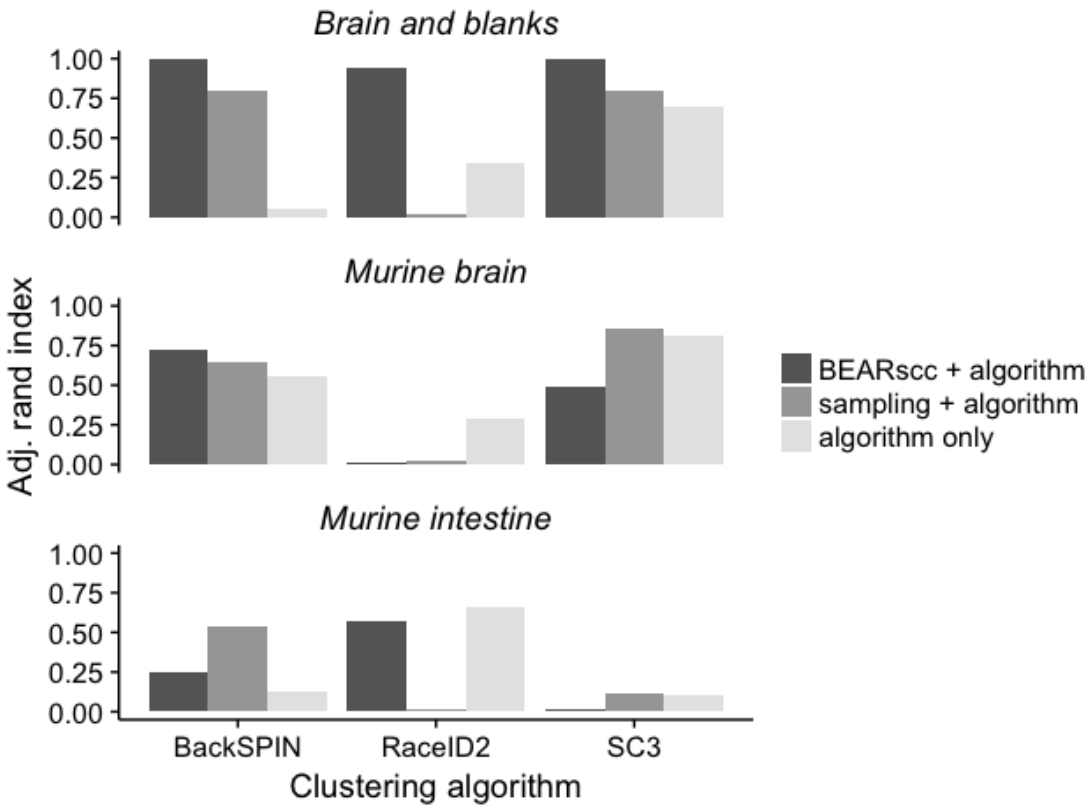


Figure 2.10.1: BEARscc was benchmarked in comparison to sampling and naive clustering with each respective clustering algorithm. Bar plots displays the adjusted rand index (y-axis) for each clustering algorithm (x-axis), where either BEARscc, gene sampling, or the clustering algorithm alone are displayed (denoted by greyscale). Each dataset is represented in a panel, where the brain and blanks control (top), the murine brain (middle) and the murine intestinal (bottom) data were analyzed.

2.11 Conclusions and discussion

To address the challenges posed by the intrinsically high technical variability of single-cell transcriptome sequencing, I developed BEARscc, a new tool that enables the evaluation of single cell clustering results with respect to technical noise. BEARscc is not a clustering algorithm in itself, but rather a tool to evaluate the results produced by third-party clustering algorithms. To do so, it aggregates the information from exogenous control spike-ins across samples to create a model of both the expected variance of endogenous read counts as well as the likelihood of false-negative measurements. This represents a novel and alternative use of spike ins controls that is not subject to the same potential issues surrounding the use of spike ins for per-sample normalization. BEARscc also provides a number of utility functions to interpret the resulting set of clusterings, ranging from visual inspection via the noise-consensus matrix to several summary statistics that indicate the robustness of individual clusters to technical noise. The source code and documentation for BEARscc are freely available from https://bitbucket.org/bsblabludwig/bearscc.

The BEARscc results concerning control as well as on real biological data show that BEARscc's model is able to produce simulated technical variation that closely resembles that observed experimentally, both when considering gene expression variance (Figure 2.5.1a,b) as well as false-negatives (Figure 2.5.1c). I cannot rule out, however, that extreme batch effects with a multi-modal distribution of variance could skew BEARscc's noise model and lead to biased simulated replicates. I envision that future versions of BEARscc will attempt to detect and warn about such biases. Furthermore, while the drop-out model calculated by BEARscc is accurate for genes with an average expression of more than one count, there is still scope for improvement. Future work will focus on more precise models of drop-outs in the context of very low gene expression.

I tested BEARscc in conjunction with three popular clustering tools (SC3, BackSPIN, and RaceID2) and showed that all three methods when used alone report false-positive clusters that are not robust to simulated technical variation (Figure 2.7.1 and 2.7.2). BEARscc flagged false-positives where the confounding factor

was library size, and also highlighted batch-induced clusters. I also benchmarked BEARscc against analogous application of a simple sampling method and found BEARscc to be superior to simple sampling in most combinations of clustering methods and datasets. Importantly, BEARscc is not designed to provide another clustering method, but to enable the user to understand their data in terms of experimental precision in an organic manner. Batch effects represent a particular challenge in scRNA-seq experiments, which is compounded by the logistical difficulty of eliminating such effects through experimental design alone (Hicks *et al.* 2015; Wagner *et al.* 2016) BEARscc's ability to identify batch-induced false positives will therefore be of particular importance as scRNAseq experiments are performed on increasing numbers of cells.

Numquam ponenda est pluralitas sine necessitate.

Plurality must never be posited without necessity.

— William of Ockham in *Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi*

3

Single cell RNA-sequencing reveals insights in Barrett's esophagus

Contents

3.1 Introduction **76**
3.2 Experimental design **79**
3.2.1 Endoscopic biopsies of four upper GI tract tissues are collected from six patients 80
3.2.2 Endoscopic biopsies are FACS sorted into single cells . . 81
3.2.3 Controls for batch and successful sequencing of transcriptomes 81
3.2.4 Description of samples sequenced with single cell RNA-sequencing 81
3.3 Data and methods **84**
3.3.1 Patient enrollment and endoscopic biopsies sampling . . . 84
3.3.2 Isolation of single cells from endoscopic biopsies 85
3.3.3 Single cell RNA-sequencing 86
3.3.4 Bulk RNA-sequencing 86
3.3.5 Library normalization 86
3.3.6 Differential expression analysis to identify marker genes . 87
3.3.7 Immunohistochemistry and immunofluorescence staining 87
3.4 Duplicate removal does not significantly alter biological signal . **88**
3.4.1 Duplication is high in our dataset 89
3.4.2 Duplication rates do not correlate with library complexity 90
3.4.3 Duplicate removal has a small impact on most genes . . . 93
3.4.4 Gene-wise duplication rates are independent of gene length and disproportionately high in highly expressed genes . . 94
3.4.5 Duplicate removal does not affect biological signal 96
3.4.6 Duplicates were removed before further analysis of our data 101

3.5 5' or 3' bias was not present in our data. **101**
3.6 Mapping metrics and QC of single cell RNA-sequencing libraries **102**
3.6.1 Mapping metrics are not effective indicators of library quality in our data 102
3.6.2 ERCC spike-in concentration in our samples 106
3.7 Mitochondrial and rRNA expression in our samples **111**
3.7.1 Nuclear rRNA gene expression is low in our samples . . . 111
3.7.2 Mitochondrial gene expression is high in our libraries . . 114
3.8 High quality libraries were obtained from endoscopic biopsies. . **117**
3.8.1 Transcriptome fragment count determines high quality libraries 117
3.8.2 Transcriptome fragment count and library complexity . . 120
3.8.3 ERCC content is higher in low transcriptome fragment libraries 125
3.8.4 Feature selection for clustering analysis 126
3.8.5 1821 samples have high quality transcriptomes 127
3.9 Overview of BEARscc analysis of Barrett's experiment libraries **129**
3.10 Determining clusters for biological interpretation **131**
3.10.1 Initial SC3 cluster number parameters in individual tissues 131
3.10.2 Initial SC3 cluster number parameter across tissue types 136
3.10.3 Final clusters are determined by BEARscc 137
3.11 BEARscc identifies substructure in individual gastro-intestinal epithelia **143**
3.11.1 Summary of subcellular structure in upper GI epithelia . 144
3.11.2 Understanding the observed cellular structure 145
3.12 Interpreting the relationships between cells of different tissue types. **147**
3.12.1 The biology of the 12 BEARscc curated clusters 148
3.12.2 Barrett's esophagus is functionally similar to esophageal submucosal glandular duct cells 154
3.12.3 OLFM4 and LEFTY1 are expressed in Barrett's esophagus 158
3.12.4 ITLN1 and SPINK4 genes mark early goblet cells 160
3.13 Conclusions and discussion **162**

3.1 Introduction

The origin of Barrett's esophagus is still under considerable debate. Barrett's esophagus was, for many years, thought to somehow originate from the normal squamous tissue of the esophagus (Hu *et al.* 2007; Kazumori 2006; Salo *et al.* 1996). More recently, a number of other hypotheses have been suggested. Among them, that Barrett's esophagus arises from bone marrow stem cells circulating in the blood (Hutchinson *et al.* 2010), from a residual embryonic stem cell subpopulation (Wang

et al. 2011), or from stem cells in the gastric cardia (Quante *et al.* 2012). To date, the prevailing theory is that gastric progenitor cells migrate into the lower esophagus and acquire potential for intestinal differentiation (Jansen & Wright 2016; McDonald *et al.* 2014; Cancer Genome Atlas Research Network *et al.* 2017). However, many of the proposed hypotheses are rooted in observations in rodent models. Importantly, the gastrointestinal tract of the rodent is anatomically distinct from humans in two critical ways: esophageal submucosal gland ducts are present in humans, but not rodents, and the histological junction of squamous and columnar tissue occurs at the gastro-esophageal anatomic junction in humans, but instead occurs in middle of the stomach for rodents (Garman *et al.* 2012). These fundamental differences in anatomy and histology and the fact that neither mice nor rats develop Barrett's esophagus under normal conditions suggest that metaplasia is initiated by different methods in rodent models (Garman *et al.* 2012).

The absence of esophageal submucosal gland ducts is particularly important when considering the utility of rodent models for Barrett's esophagus (Garman *et al.* 2012). These gland ducts, more commonly found in the distal esophagus of humans, may provide a protective role for the squamous esophagus in humans as the ducts are known to secrete alkaline fluid, thereby neutralizing acidic contents in the esophageal lumen (Long & Orlando 1999; Savarino *et al.* 2016). Intriguingly, it has been suggested that esophageal submucosal gland ducts are a developmental "remnant" of the glandular epithelium preceding the squamous epithelium. In fact, it has even been proposed that multi-potent stem cells found in these gland ducts could be the origin of Barrett's esophagus (Lörinc *et al.* 2015). The clinical observation that hyperplastic esophageal submucosal gland ducts are more prevelant beneath Barrett's esophagus columnar epithelia (van Nieuwenhove *et al.* 2001; Lörinc & Öberg 2015), and the fact that these gland ducts share many structural and protein expression features with gastric glands (McDonald *et al.* 2014), further implicates the esophageal submucosal gland ducts in the origination of Barrett's esophagus. However, no conclusive evidence identifies these gland ducts as the progenitor for Barrett's esophagus, and there is controversy whether gastric or esophageal sub

mucosal glands, if either, contribute to the onset of Barrett's esophagus (McDonald *et al.* 2014; van Nieuwenhove *et al.* 2001; Lörinc & Öberg 2015).

We proposed to identify the functional relationship between Barrett's esophagus and the three normal tissues with histological features found in Barrett's (i.e. the duodenum, gastric tissue, and the proximal normal squamous tissue) using single cell RNA-sequencing. Prior work by others successfully distinguished Barrett's esophagus from other upper gastrointestinal stem cell populations (Yamamoto *et al.* 2016) and recent work has successfully identified rare cell subpopulations using single-cell transcriptomics (Grün *et al.* 2015; Grün *et al.* 2016b; Ennen *et al.* 2014). To that end we have performed single cell RNA-sequencing and subsequent analysis on 3043 samples from controls, the Barrett's lesion, normal gastric epithelia, normal duodenal epithelia, and normal proximal squamous esophageal epithelia from four patients presenting with Barrett's esophagus, and normal squamous esophageal epithelea from two patients presenting with gastric reflux and no evidence of Barrett's esophagus.

In the following work, I describe the design of our experiment (Section 3.2), general data and methods (Section 3.3), quality control, and biological conclusions drawn from applying BEARscc in conjunction with SC3 (Kiselev *et al.* 2016) to the 3051 samples. During quality control analysis, I determine that exact position duplicates should be removed (Section 3.4), and I find that both 5 prime and 3 prime biases were absent in our data (Section 3.5). Furthermore, I demonstrate that rRNA expression is universally low in biological samples, while mitochondrial expression is a large component of all libraries, including whole tissue samples from the same patient (Section 3.7). Although I observe generally poor mapping metrics in our data (Section 3.6), I conclude that high quality transcriptomes were obtained from endoscopic biopsies relying primarily on a threshold of total fragments mapped to the transcript (Section 3.8). This threshold was primarily informed by the positive and negative control samples. I then present an overview of downstream analysis performed by BEARscc in order to define subpopulations within and across tissue types and infer biology from the single-cell libraries(Section 3.9). After determining the number of clusters for each tissue type and across tissue

types (Section 3.10), I describe the functional architecture of the three normal tissue types (proximal squamous esophageal, gastric, and duodenal tissue), and the Barrett's lesion (Section 3.11). I similarly characterize clusters formed from all QC-filtered cells take together, and examine the relationships between cells across tissue types (Section 3.12). Moreover, by exploring the relationship across tissue types, I demonstrate, with supporting work from Dr. Richard Owen, that Barrett's esophagus tissue has profound similarity to esophageal submucosal glandular ducts (Section 3.12.2). Finally, new marker genes for early goblet cells (Section 3.12.4) across upper gastro-intestinal tissue types and stem-like subpopulations (Section 3.12.3) in Barrett's and esophageal submucosal glandular ducts identified in the single cell RNA-sequencing data by myself and Dr. Richard Owen are validated *ex vivo* by Dr. Michael White.

3.2 Experimental design

Barrett's esophagus is histologically diverse with normal squamous esophageal, gastric, and intestine phenotypic character. In order to shed light on the origins of this abbarent histology of the lower esophagus, we aimed to characterize the histological heterogeneity found in Barrett's. Specifically, we wanted to understand the cellular substructure of the normal squamous esophagus, gastric cardia, and the intestine controls in order to describe the tissue heterogeneity of Barrett's at the molecular level and potentially identify the functional tissue type of origin. In order to do this, we sequenced 2095 cells from endoscopic biopsies collected from the upper GI tract of six patients and performed whole tissue RNA-sequencing on proximal tissue from endoscopic biopsy samples of the upper GI tract for four of these patients. Drs Owen and White obtained double bite quadrantic 2mm endoscopic biopsies from gastroenterologists, Drs. Adam Bailey and Barbara Braden, for the purpose of this experiment.

3.2.1 Endoscopic biopsies of four upper GI tract tissues are collected from six patients

Dr. Richard Owen and Dr. Michael White recruited patients undergoing routine endoscopic surveillance for Barrett's esophagus and mild gastric reflux symptoms (see Methods 3.3.1). Six patients were scheduled to undergo gastroscopy and were consented to provide endoscopic biopsy samples (REC:09/H060606/5+5).

Four of the selected patients (GEN02021, GEN02023, GEN02024, GEN02025) had presented with Barrett's esophagus. In order for this work to be applicable internationally, Barrett's esophagus was more stringently defined by the American definition which requires the presence of intestinal metaplasia in addition to the columnar phenotype. From these patients, endoscopic biopsy samples were taken from the second part of the duodenum, the gastric cardiac region 20mm distal to the gastro-esophageal junction, and the proximal normal esophageal squamous epithelium at least 20mm distal to the closest Barrett's tissue (Figure 3.2.1a,c). 2-4 2mm endoscopic biopsies were taken for each control tissue in each patient. Additionally, 2-4 2mm biopsies were also taken from a central region of the Barrett's lesion in each patient; care was taken to ensure that for each sample, the gastro-esophageal junction and proximal Barrett's esophagus margin were avoided. Every biopsy was cut into thirds, where one third was pooled by patient and tissue to be FACS sorted into single cells for sequencing. Another third was pooled by patient and tissue for whole tissue RNA sequencing. The final third was pooled by patient and tissue in order to perform histology and downstream experimental validation. Importantly, the histology for each tissue and patient was verified prior to further experimentation (Supplementary Figure A.1.4).

Two of the selected patients (GEN2026, GEN02027) had presented with gastric reflux symptoms, but no Barrett's esophagus symptoms. From these patients, 2-4 2mm endoscopic biopsies of the normal esophagus were taken from each of these patients (Figure 3.2.1b-c). These biopsies were pooled by patient, disassociated, FACS sorted into single cell samples, and sequenced. These patients provide a

normal esophageal control for comparison with histologically normal esophagus observed in patients that have presented with Barrett's esophagus.

3.2.2 Endoscopic biopsies are FACS sorted into single cells

Sample fragments for single cell analysis were suspended in digestion solution and disassociated mechanically in buffer (See Methods 3.3.2). Disassociated biopsy suspensions were then sorted into single cells, where observations on the FACS were filtered for size and viability (Supplementary Figure A.1.5). Furthermore, observations on the FACS were also filtered using antibodies to remove immune cells and to select for epithelial character. Filtered cells were directly sorted into 96-well plates containing lysis buffer and immediately frozen for RNA extraction, cDNA conversion, and library preparation for sequencing (see Methods 3.3.3).

3.2.3 Controls for batch and successful sequencing of transcriptomes

Logistically, sorting disassociated primary endoscopic biopsies from patients with FACS into single cells often results in 96-well plates that are completely confounded by batch. To prevent results from being completely confounded by batch, we re-pipetted cells from the original 8 frozen 96-well plates that formed batches B and C into a new set of 8 96-well plates, each of the four tissues and multiple patients were represented on each plate. Additionally, at least two wells on each plate were reserved for positive and negative controls. Positive controls consisted of dilute whole tissue brain samples (see Methods 2.2.1) and negative controls consisted of blank water only samples (see Methods 2.2.1).

3.2.4 Description of samples sequenced with single cell RNA-sequencing

In total we sequenced 3051 samples at single cell RNA concentrations, where 156 of these were libraries from positive and negative controls and 2895 were libraries from single cells isolated from endoscopic biopsies (Supplementary Table A.2). The number of cells were relatively well distributed across patients, except

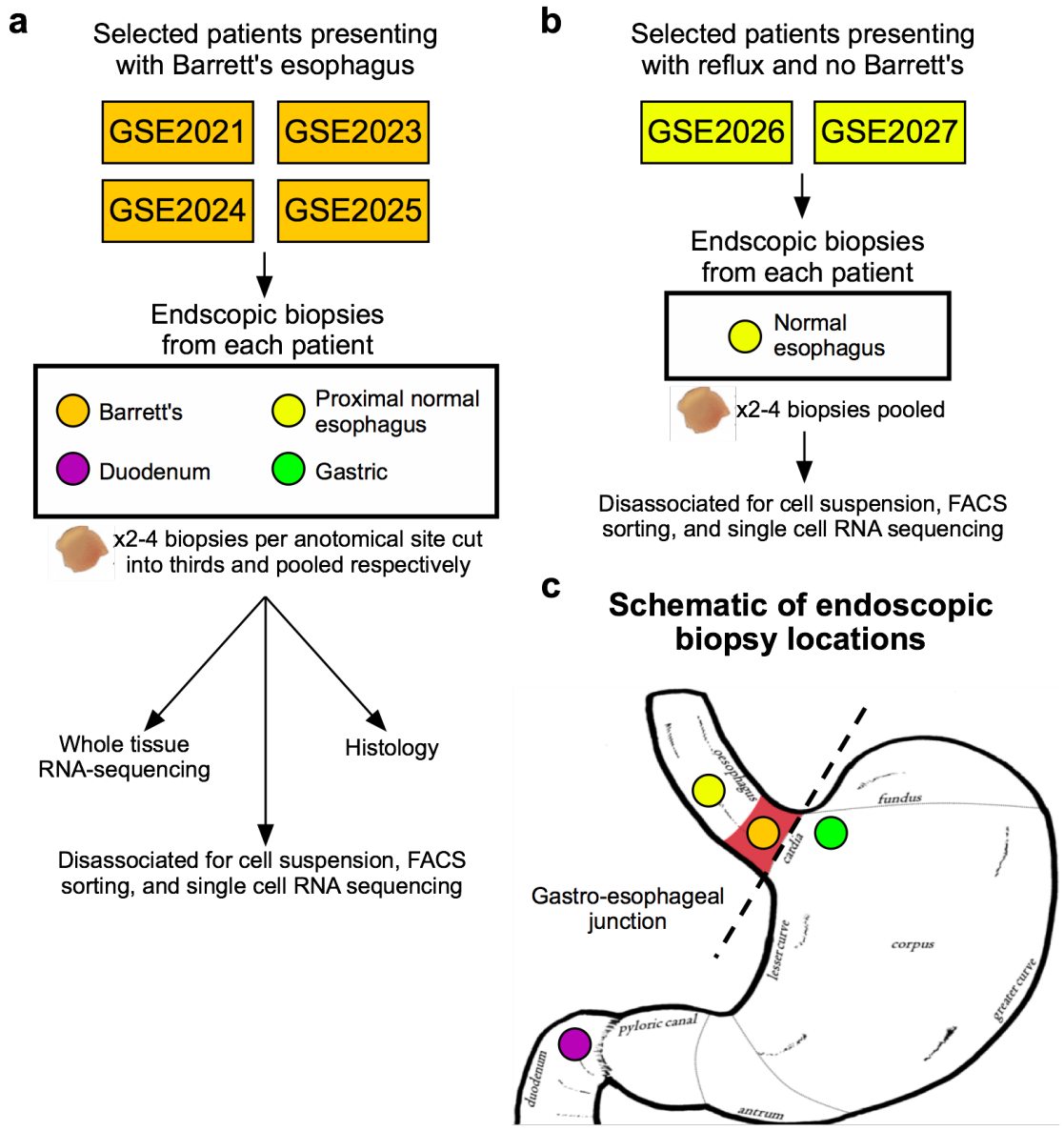


Figure 3.2.1: **a**, The schematic illustrates how for each of the four patients presenting with Barrett's esophagus, 2-4 2mm biopsies were taken from each anatomical site: Barrett's esophagus, normal proximal squamous esophagus, gastric cardia, and the duodenum. For each anatomical site the 2-4 biopsies were cut into thirds, where one third from each biopsy was pooled for either histology, whole tissue RNA-sequencing or single cell RNA-sequencing. Tissues are coded by color and can be referenced to the anatomical schematic drawn by Dr. Richard Owen in panel (**c**) and all the following tissue-related figures in this chapter. **b**, The schematic illustrates how for each of the two patients presenting with gastric reflux, but no Barrett's esophagus, 2-4 endoscopic biopsies were taken and pooled for single cell RNA-sequencing. **c**, An anatomical illustration of the human upper gastrointestinal tract, where the gastro-esophageal junction is explicitly shown by dotted lines and the sampling points for the endoscopic biopsies are indicated by color. The Barrett's lesion that would be found in patients from panel (**a**) is indicated in red.

that cells from GEN02021 were over-represented and cells from GEN02026 were under-represented in the experiment (Table 3.1).

Patient	Sample #	Clinical presentation
GEN02021	851	Barrett's esophagus
GEN02023	427	Barrett's esophagus
GEN02024	445	Barrett's esophagus
GEN02025	453	Barrett's esophagus
GEN02026	180	Gastric reflux with no Barrett's
GEN02027	539	Gastric reflux with no Barrett's
Controls	156	na

Table 3.1: Number of samples sequenced from each patient and number of control wells

Sample libraries were sequenced in 6 batches (batches A-F), where batches A-E contained libraries isolated from patients presenting with Barrett's esophagus and batch F consisted of libraries isolated from patients presenting with gastric reflux only. Tissue and patient origin for cells on plates from batches A, B, and E were somewhat confounded by batch, while batches C and D did not appear to be confounded at all as previously described (Section 3.2.3). Batch F consisted of all the samples from the two patients presenting with gastric reflux, but no evidence of Barrett's esophagus. Batches A, B, and E were smaller (approximately 260 samples), while batches C,D, and F were larger (approximately 750 cells) (Table 3.2).

Batch	Sample #	Clinical presentation
Batch A	179	Barrett's esophagus
Batch B	221	Barrett's esophagus
Batch C	768	Barrett's esophagus
Batch D	735	Barrett's esophagus
Batch E	381	Barrett's esophagus
Batch F	767	Gastric reflux with no Barrett's

Table 3.2: Number of samples sequenced in each experimental batch

From the patients presenting with Barrett's esophagus we sequenced a total

of 620 cells from the Barrett's lesion, 678 cells from the proximal gastric tissue, 526 from proximal normal squamous esophagus, and 352 cells from the duodenum (Table 3.3). These libraries isolated from endoscopic biopsies were sequenced in parallel with 61 positive (dilute whole brain tissue) and 47 negative (water) controls. From the patients presenting with gastric reflux only, we sequenced a total of 719 normal squamous esophageal cells, 61 positive (dilute whole brain tissue) and 47 negative (water) blank controls (Table 3.3).

Tissue of origin	Sample #	Clinical presentation
Barrett's lesion	620	Barrett's esophagus
Proximal squamous esophagus	526	Barrett's esophagus
Normal squamous esophagus	719	Gastric reflux with no Barrett's
Proximal gastric	678	Barrett's esophagus
Duodenum	352	Barrett's esophagus
Dilute whole brain (pos. control)	85	na
Water blanks (neg. control)	71	na

Table 3.3: Number of samples sequenced from each tissue or control

3.3 Data and methods

The following describes in detail the relevant data collection and other methods employed throughout this chapter to generate results.

3.3.1 Patient enrollment and endoscopic biopsies sampling

Patient recruitment, consent, and enrollment were completed by Drs. Richard Owen and Michael White. Patients were enrolled that were attending routine endoscopic monitoring of Barrett's esophagus or were having a gastroscopy for diagnosis following presentation with mild reflux symptoms. These patients were consented for provision of endoscopic biopsies (REC:09/H0606/5+5). Standard biopsy forceps (Radial Jaw 4 Standard Capacity, Boston Scientific, Natick, USA) were used to sample 2-4 double bit quadrantic 2mm biopsies from anatomical sites of interest. Specifically, from each patient presenting with Barrett's biopsies

were sampled from a central region of Barrett's esophagus avoiding the proximal Barrett's esophagus margin and the gastro-esophageal spectrum. Biopsies were also sampled from the gastric cardia tissue 20mm distal to the gastro-espophageal junction, the second part of the duodenum, and normal squamous esophageal epithelium at least 20mm distal to the proximal extent of the Barrett's lesion. Only normal squamous esophageal epithelium was biopsied from patients with mild reflux symptoms. Patients were selected based on previously characterized clinical and pathological features (Supplementary Table A.1).

3.3.2 Isolation of single cells from endoscopic biopsies

Isolation of single cells from endoscopic biopsies was performed by Drs. Richard Owen and Michael White. One third of each endoscopic biopsy was pooled by tissue and patient, and placed into digestion solution (1x phosphate buffered solution (Gibco^{TM}, heat inactivated fetal bovine serum (Sigma-Aldrich®, EDTA, and type 1 collagenase (Worthington Biochemical Company®). Samples were then mechanically disassociated mechanically with 60 minutes of gentle oscillation, further fragmentation with scissors, and briefly manually pipetted up and down with a p1000. Large fragments of tissue were alllowed to settle. The supernatant containing the cell suspension was filtered (100 micron Sysmex Celltrics®) into a 15mL Falcon tube. The process was repeated 3 times, where the final supernatant was centrifuged. The resulting cell pellet was resuspended in sorting buffer, and a small amount of all samples were pooled for labelling controls. In preperation for FACS, pre-conjugated CD45-FITC (immune selection) and EpCAM-PE(epithelial selection) antibodies were added into cell suspensions. These were incubated and washed according to the manufacturer's advice. To select for viability, immediately prior to sorting, DAPI (Sigma-Aldrich®) was added to cell suspensions. Cell were selected in FACS (BD Biosiences FACS Ariallu platform - 70 μm nozzle) based on size and singlet gating in order to saturate cell output and minimize debris in subsequent gates. Then, low CD45+ and high EpCAM+ cells were selected to enrich for epithelial cells for downstream sequencing. Resulting cells were directly sorted into 96-well

plates (Life Technologies™ MicroAmp®Optical 96-well Reaction Plate), where each well was prepared with 2μL 0.2% Triton™ X-100 and RNAse inhibitor (Takara Recombinant RNase Inhibitor) at 19:1. Once filled, a plate was then immediately frozen on dry ice. Spectral accuracy was confirmed with compensation bead controls and pooled cell suspensions as fluorescence-minus-one controls. In all cases, the process was limited to 3hours from the time of endoscopic biopsy extraction until 96-well plates containing lysed single cell suspensions were frozen on dry ice.

3.3.3 Single cell RNA-sequencing

Transcriptome libraries from endoscopic biopsies were diluted to 10pg aliquots and added to $1\mu L$. cDNA conversion, library preparation, and sequencing were performed by the Wellcome Trust Center for Human Genomics Sequencing Core. Samples were pipetted into 96-well plates and treated as single cells using Smartseq2 cDNA conversion as described by Picelli *et al.* (Picelli *et al.* 2013) with minor modifications. The library was prepared using Fludigm's recommendations for Illumina NexteraXT at $\frac{1}{4}$ volume with minor modifcations, and sequenced on the Illumina HiSeq4000 platform. Raw reads were mapped to hg19 with STAR(Dobin *et al.* 2013). Exact position duplicates were filtered, and fragment counts for genes were computed with HTseq (Anders *et al.* 2015). Positive and negative controls were prepared as previously described (see Section 2.2.1).

3.3.4 Bulk RNA-sequencing

Raw reads were mapped to hg19 and counted using STAR(Dobin *et al.* 2013).

3.3.5 Library normalization

For expression measurements reported in transcripts per million (TPM), an adjusted library size was calculated for each library using trimmed mean of M-values (TMM) normalization method with the edgeR package (Robinson *et al.* 2009). The expression values were then normalized by gene length and the resulting

adjusted library size. In all cases where count values are directly presented, TPM values are used unless otherwise stated. This was to ensure that observations were comparable across both cells and genes. In contrast, for analysis by BEARscc, libraries were normalized as suggest by Hicks *et al.* and described in Section 2.6. As is best practice, counts were not normalized prior to differential expression analysis with edgeR (Robinson *et al.* 2009).

3.3.6 Differential expression analysis to identify marker genes

The marker genes for clusters was determined using standard differential expression analysis with edgeR (Robinson *et al.* 2009). Because differential expression analysis assumes that most genes are not differentially expressed, which may not be true in single-cell analysis, the null distribution of differential expression false discovery rate was computed empirically. Specifically, identical differential expression analysis was performed 1000 times, where membership of the candidate cluster was scrambled with members of other clusters. The most significant p-value was taken from each iteration to create a null distribution of random p-values specific to the each differential expression experiment. The false discovery rate (FDR) for a given differential expression result in the test for marker genes was taken to be the percentile of the observed p-value on the empirically derived null p-value distribution.

3.3.7 Immunohistochemistry and immunofluorescence staining

All immunohistochemistry and immunofluorescence work was performed by Dr. Michael White. Endoscopic mucosal resection specimens form 35 patients with Barrett's esophagus and normal squamous esophageal samples from 5 patients' oesophagectomies were obtained from the Oxford biobank. These sections were de-waxed, rehydrated and then incubated for 10 minutes in 3% hydrogen peroxide in methanol in order to block endogenous peroxidase activity. pH6 sodium citrate was applied to sections for antigen retrieval. After normal goat serum was used to block sections, they were incubated overnight at 4°C with primary

antibody against SPINK4 (1:2000, rabbit polyclonal, Abcam), ITLN1 (1:2000, rabbit polyclonal, Abcam), MUC2 (1:500, rabbit monoclonal, Abcam), CHGA (1:1000, mouse monoclonal, Abcam), LEFTY1 (1:1000), Ki67 (1:1000), KRT14 (1:1000, Abcam), KRT8 (1:1000), TFF3 (1:1000, mouse monoclonal, Sigma Aldrich). For immunohistochemical staining, the samples were treated with biotinylated secondary antibody (Vecor Labs, 1:250) for 40 minutes at room temperature. The staining reaction was worked up (Vector Elite ABV kit) and counterstained with haemotoxylin. Pathologists, Drs. Lai Mun Wang and Rob Goldin, examined all samples under a histology microscope. For immunofluorescence staining, the expression was detected (Alexa Fluor, 1:250, Molecular Probes for 1 hour). Nucleic acids were stained with DAPI (1:2000), and samples were observed under a confocal microscope (LSM 710; Carl Zeiss).

3.4 Duplicate removal does not significantly alter biological signal

After obtaining raw counts for each cell in batch A and B, I determined to investigate the effects of duplicate removal in terms of technical and biological signals. To that end, I filtered all single cell library BAM files to remove all identical fragments as determined by mapped position (same 5' read position and 3' read position on the reference) and used HTseq (Anders *et al.* 2015) to determine raw counts for transcriptome features. I repeated this analysis with unfiltered BAM files for all single cell libraries. The work that follows serves as an assessment of the nature of duplicates in single cell RNA-seq experiments. Importantly, the work that follows includes cell line samples (OE33) and upper GI samples from patients that were excluded from our final cohort, because they were part of pilot experiments that informed the experimental design. These samples were processed, sorted and sequenced in the same way as the samples from the final. A number of the results reported in this section are expressed in Δ, which is a measure of the change in reported gene expression for samples when duplicates are removed vs. retained. Specifically, if x is the reported gene expression for a gene with

duplicates retained, and y is the reported gene expression for a gene with duplicates removed. Then, the fractional change in reported gene expression due to duplicate removal, or Δ, can be described by equation 3.1.

$$\Delta = \frac{x - y}{x} \cdot 100 \tag{3.1}$$

I utilize Δ frequently in this section in order to understand the effects of duplicate removal on reported gene expression.

3.4.1 Duplication is high in our dataset

In order to determine whether or not to maintain position duplicates within our single cell RNA-sequencing samples, I first examined whether duplicates represented a significant issue. Since single cell libraries consist of extremely low RNA content starting material by definition, it was not surprising that I found duplication to be high in our experiment. Indeed, I found duplication to be extraordinarily high with the vast majority of libraries composed of at least 60% duplicate alignments (Figure 3.4.1). Therefore, I felt it was important to understand the impact of duplicates on both technical and biological signals.

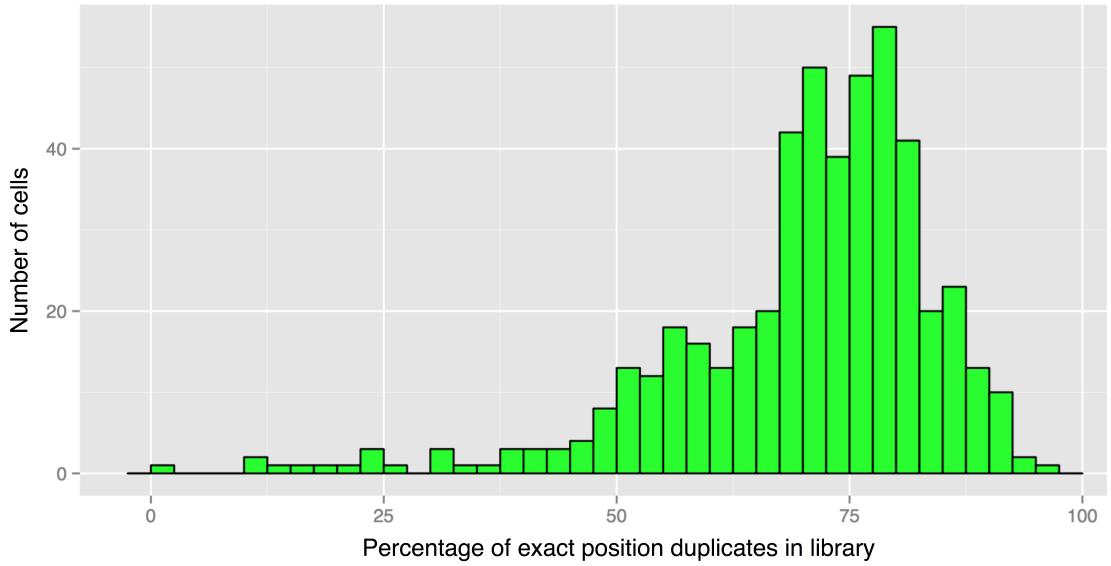


Figure 3.4.1: The number of duplicates in our single cell RNA-sequencing samples is high. Histogram illustrates the percentage of alignments that are exact position duplicate alignments across single cell libraries. Each bar represents the number of single cell libraries with a duplication rate within a binned percentage of duplications.

3.4.2 Duplication rates do not correlate with library complexity

I reasoned that high duplication rates were a result of samples with low library complexity. That is, I suspected that duplication rates would increase dramatically in samples that had less diverse starting RNA material for conversion and amplification. However, I found that in single cell libraries and dilute controls (whole tissue brain, water, and OE33 cell line samples), the duplication rate was not correlated with library complexity in any meaningful way (Figure 3.4.2). Additionally, aside from blank samples there were no obvious associations between the samples' origin and the the library complexity or duplication rate (Figure 3.4.2). However, blank samples exhibited low library complexity, which was expected. Finally there was no obvious relationship between the sequencing batch and duplication rate or library complexity (Figure 3.4.3).

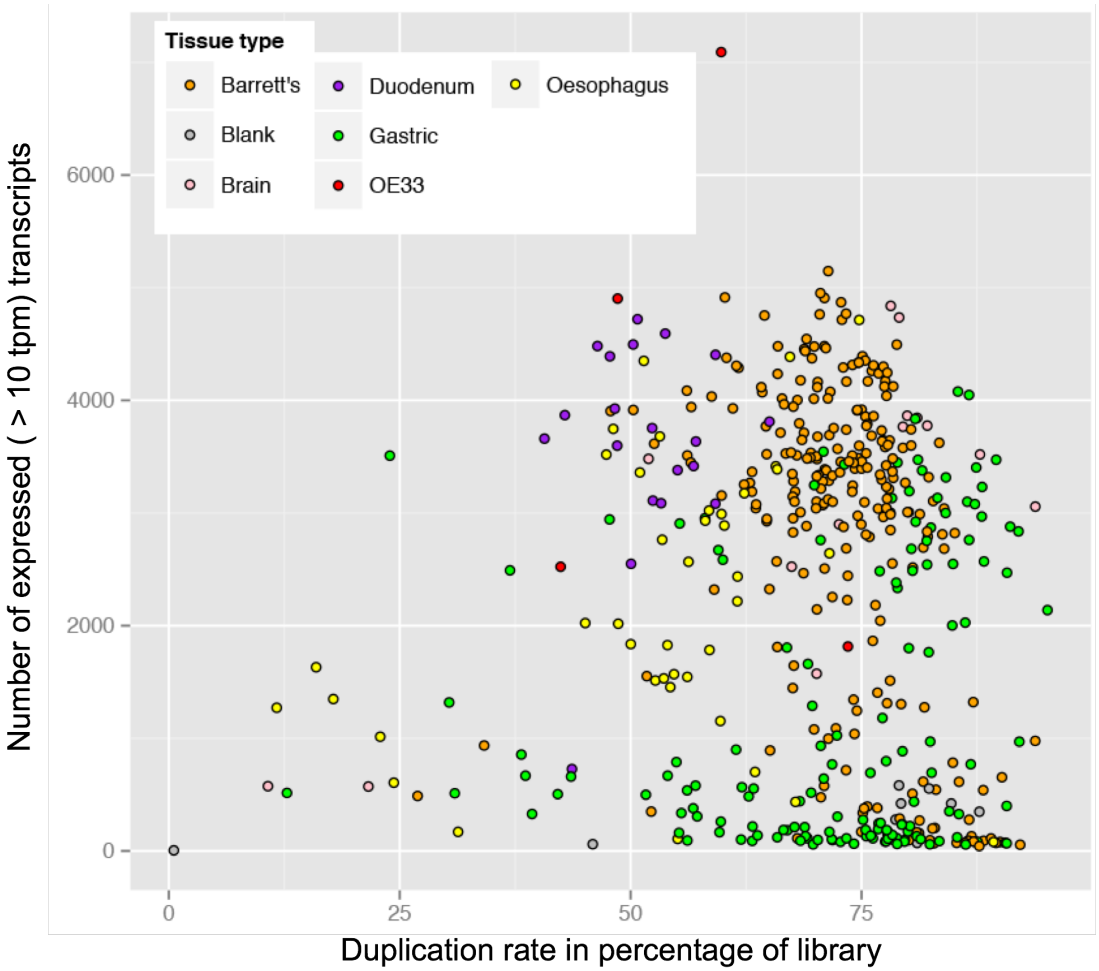


Figure 3.4.2: Scatterplot displays the number of genes with at least 10 TPM expression level as a function of the fraction of alignments that were exact position duplicate alignments. Each point is a sample and is colored by the tissue, cell line, or control sample of origin.

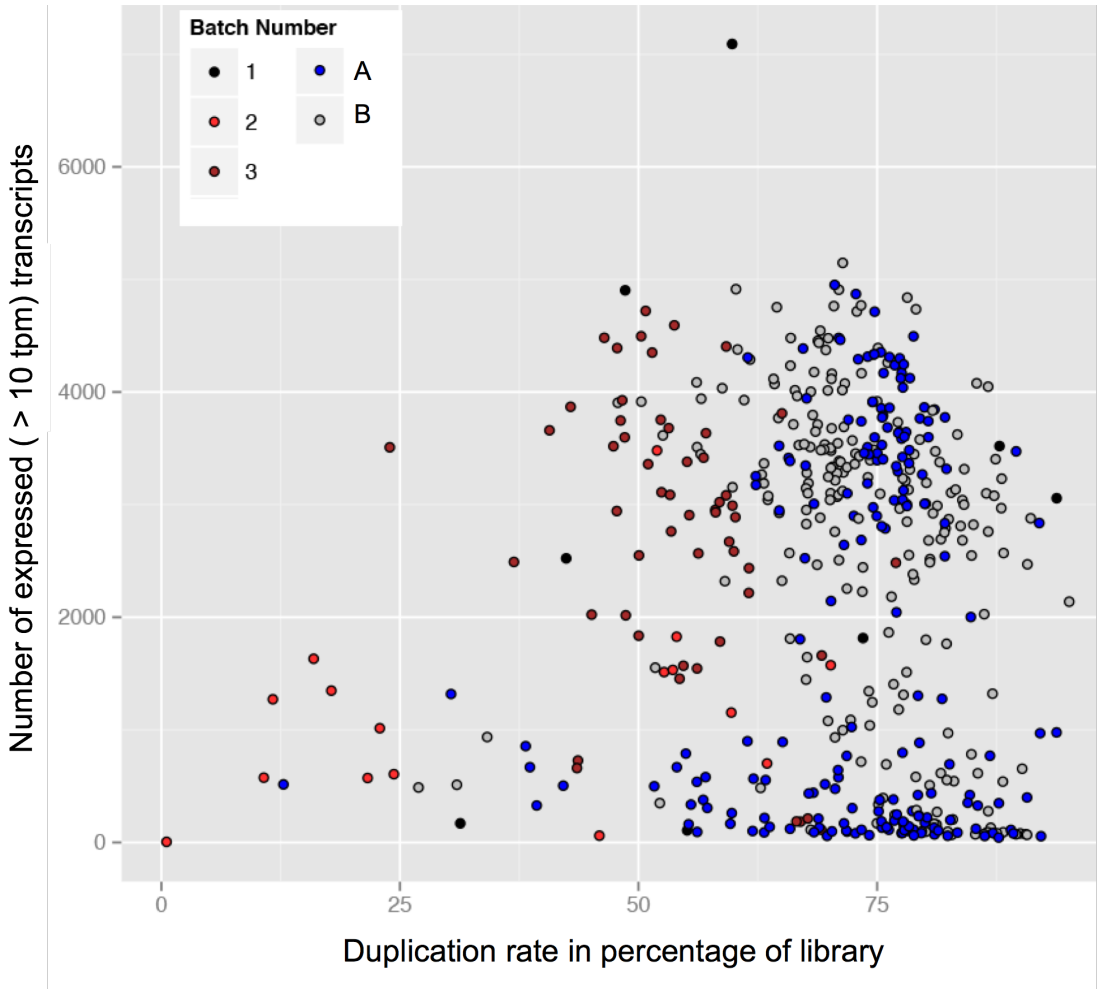


Figure 3.4.3: Scatterplot displays the number of genes with at least 10 TPM expression level as a function of the fraction of alignments that were exact position duplicate alignments. Each point is a sample and is colored by the library preparation and sequencing batch in which it was analyzed.

3.4.3 Duplicate removal has a small impact on most genes

Since I had observed extremely high duplication rates, but library complexity did not correlate with duplication rate, I investigated the effect of removing duplicates on the quantified gene expression in raw counts. To that end, I compared the measured expression level of genes (in raw counts) in samples with duplicates retained and the expression level (in raw counts) of the same samples with duplicates removed on a gene by gene basis. I observed that most genes are unaffected by duplicate removal (Figure 3.4.4). In fact, duplicate removal changes gene expression by less than 5% for virtually all genes. This result left us somewhat confused as the fraction of alignments that were duplicates was extremely large, yet most genes were unaffected by duplicate removal. Intriguingly, around 1500 cells actually increased by 1-3% when duplicates were removed. I suspect that these small changes are a result of the removal of multi-mapping reads on other sites of the genome, thereby resulting in HTseq counting reads on genetic features that were initially ignored due to multiple mapping sites.

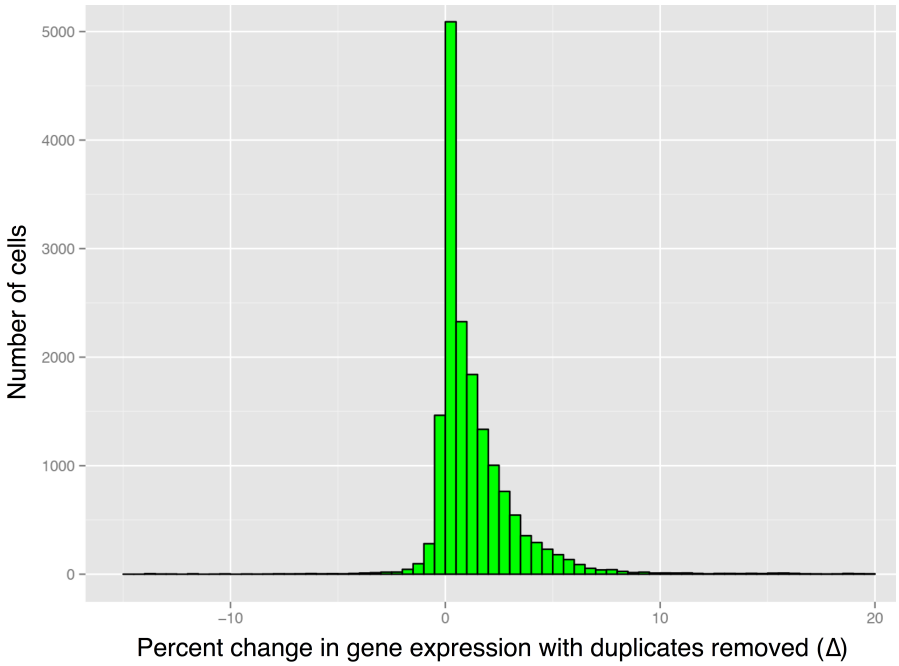


Figure 3.4.4: Reported gene expression values for each gene were computed for BAM files with duplicates removed and duplicates retained. Histogram illustrates the distribution of Δ reported gene expression as a result of duplicate removal.

3.4.4 Gene-wise duplication rates are independent of gene length and disproportionately high in highly expressed genes

To understand the discrepancy between the low number of genes with extensive changes in gene expression upon duplicate removal and the high rate of duplicate alignments, I investigated which genes were impacted by duplicate removal. Intuitively, in the event duplicates represented genuine biology, short genes would have inherently high numbers of duplicates. This intuition is a major motivator for duplicate retention in whole tissue RNA-sequencing data; highly expressed genes with short transcripts are likely to produce identical alignments following fragmentation and sequencing. To that end, I examined the change in gene expression measurements (in raw counts) between samples with duplicates removed and retained, Δ, in three bins of transcript length: short (0-1kb), medium (1-3kb), long (3-500kb) (Figure 3.4.5). I did not see any obvious differences in reported expression as a function of gene length suggesting that duplicate reads in our experiment may not reflect biological features in our single cell samples.

Next I hypothesized that, instead of representing short transcripts at high concentrations with a high probability of read sampling, the high duplication rates could be purely the result of highly sampled transcripts from highly expressed genes. I reasoned that highly expressed transcripts would dominate the cDNA pool prior to amplification by PCR. Thus, early cycles of PCR amplification would be more likely to amplify fragments from the most highly expressed genes in the library. This would suggest that the duplicate reads present in our libraries reflect PCR amplification bias rather than real biological differences in samples.

In order to test this hypothesis I examined Δ for genes of various levels of reported expression when duplicates are retained. Specifically, I examine the distribution of Δ for genes, where genes were split into three expression level groups: extremely low expression (0-5 raw counts), low expression (6-25 raw counts), and high expression (26 counts and higher) (Figure 3.4.6). I observed the distributions of Δ drastically shifted right as gene expression levels increased. That is, at

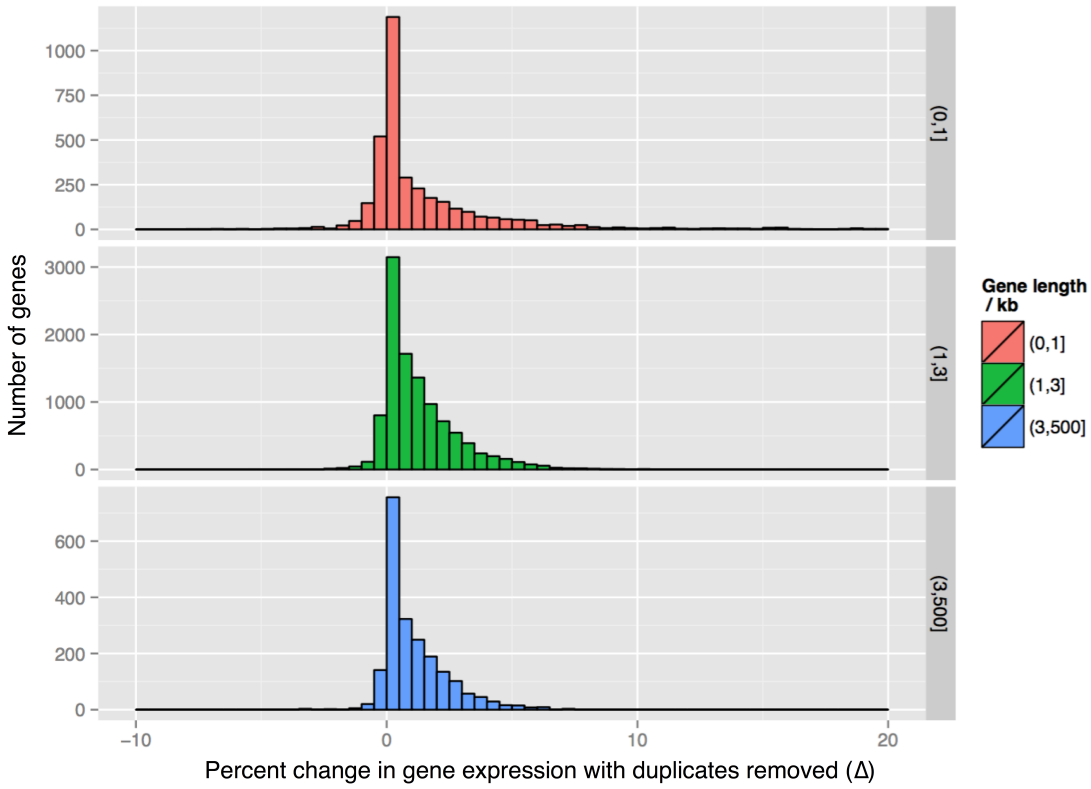


Figure 3.4.5: Reported gene expression values for each gene were computed for BAM files with duplicates removed and duplicates retained. Histogram illustrates the distribution of Δ reported gene expression as a result of duplicate removal across three different bins of transcript length, which are indicate by panel and color.

higher gene expression levels, many more genes had large changes in reported gene expression when duplicates were removed compared to when they were retained. This suggested that duplicates represented an increase in sequencing observations that corresponded directly to transcript concentration in the sample.

Taken together with the observation that the distribution of Δ across genes was unchanged for genes of different lengths, this observation suggests that identical positional duplicates did not represent a genuine sampling of highly expressed short transcripts, but were likely PCR amplification artifacts that occurred as a result of the low RNA concentrations inherent to single cell RNA-sequencing. Before electing to remove duplicates I wanted to understand the impact of duplicate retention and removal on the biological signal in the experiment.

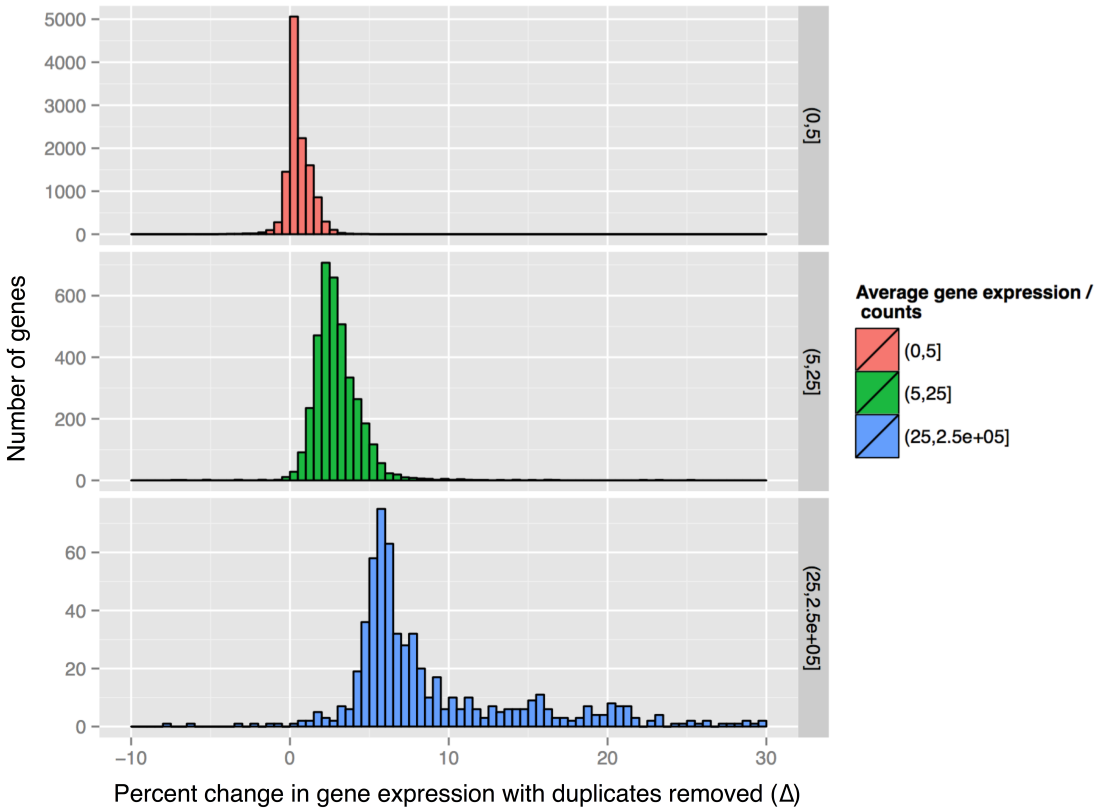


Figure 3.4.6: Reported gene expression values for each gene were computed for BAM files with duplicates removed and duplicates retained. Histogram illustrates the distribution of Δ reported gene expression as a result of duplicate removal across three different bins of reported gene expression when duplicates are retained, which are indicate by panel and color.

3.4.5 Duplicate removal does not affect biological signal

In order to understand the impact of duplicate retention and removal on the biological signal of the experiment, I examined the data in a two-dimensional projection using principal component analysis with and without duplicates (Figures 3.4.7 and 3.4.8). I reasoned that this approach would be effective due to the well-controlled nature of the experiment. I knew that the squamous epithelium of the esophagus was the most histologically distinct tissue of three controls, where functional differences between gastric and intestinal columnar tissue should be the second most biologically distinct feature of the tissue. Furthermore, I hypothesized that cells from the columnar Barrett's esophagus should share features with both gastric and intestinal tissue types. Thus, with multiple tissue type samples from

a single patient I had a clear biological expectation for the data.

Upon examining the first two (Figure 3.4.7a - left) dimensions and the second and third dimension (Figure 3.4.7a - right) of the projection for cells from control samples with duplicates retained, I noticed that, as expected, cells from control samples clearly separated by tissue type. Furthermore, the second dimension definitively separated cells of columnar origin form cells of squamous origin and the third dimension separated to some extent the cells of gastric and duodenal origin. I projected the cells form Barrett's esophagus onto the aforementioned dimensions and observed that these cells existed as a functional and columnar-epithelial smear in the projections (Figures 3.4.7b).

As a side observation, the variance explained by each dimension of the principal component analysis was universally low with the first dimension only explaining 5% of the variance (Figure 3.4.7c). I suspect this to be a result of the high level of noise inherent to single cell RNA-sequencing. Intriguingly, the first principal component was not defined by cell cycle as observed in other single cell experiments. This discrepancy between our experiment and previous work may be a consequence of our unique experimental design. The biological signal distinguishing cells sampled from multiple and functionally distinct tissue types, as was done in this experiment, may be stronger than the signal defining cells sampled from a single macro-population (e.g. t-cells or a cell line).

I performed an identical analysis with duplicates removed from the mapped reads before quantifying transcription (Figure 3.4.8). The two-dimensional projects were nearly identical for both control samples Figure (3.4.8a) and control samples with the Barrett's lesion (Figure 3.4.8b). Most certainly, the cells' tissue-specific signal was preserved in all four projections. The variance explained by the first two dimensions did increase slightly, but remained low (Figure 3.4.8c). These analyses suggest that the majority of the variance in our data was the result of the core biological signal, where tissue from different sites taken from the same patients have distinct features on the transcriptome. Furthermore, the decision to retain or remove duplicates did not distort this signal in any striking way. If anything,

duplicate removal resulted in the first two dimensions explaining more variance, where the resulting projection clearly separated cells in a biologically intuitive way.

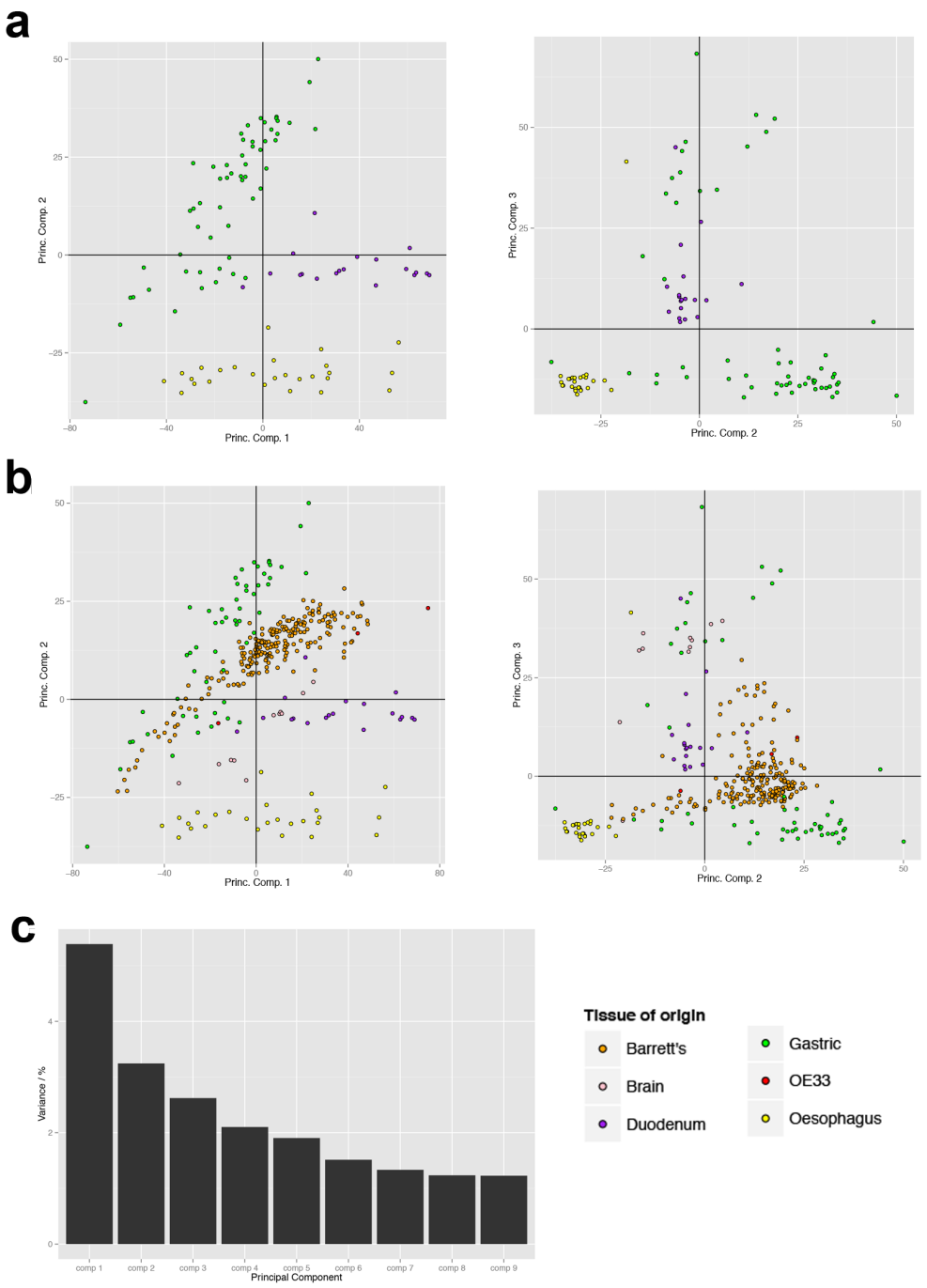


Figure 3.4.7: Tissue type is the dominant signal in the single cell libraries when duplicates are retained. Principal component analysis dimensions 1,2, and 3 of preliminary single cell data with duplicates retained. **a**, Control tissue types informed the vector space rotation and **b** Barrett's samples were treated as supplementary variables and projected onto the vector space in **a**. **c**, The variance explained by each component and the legend for the upper panels.

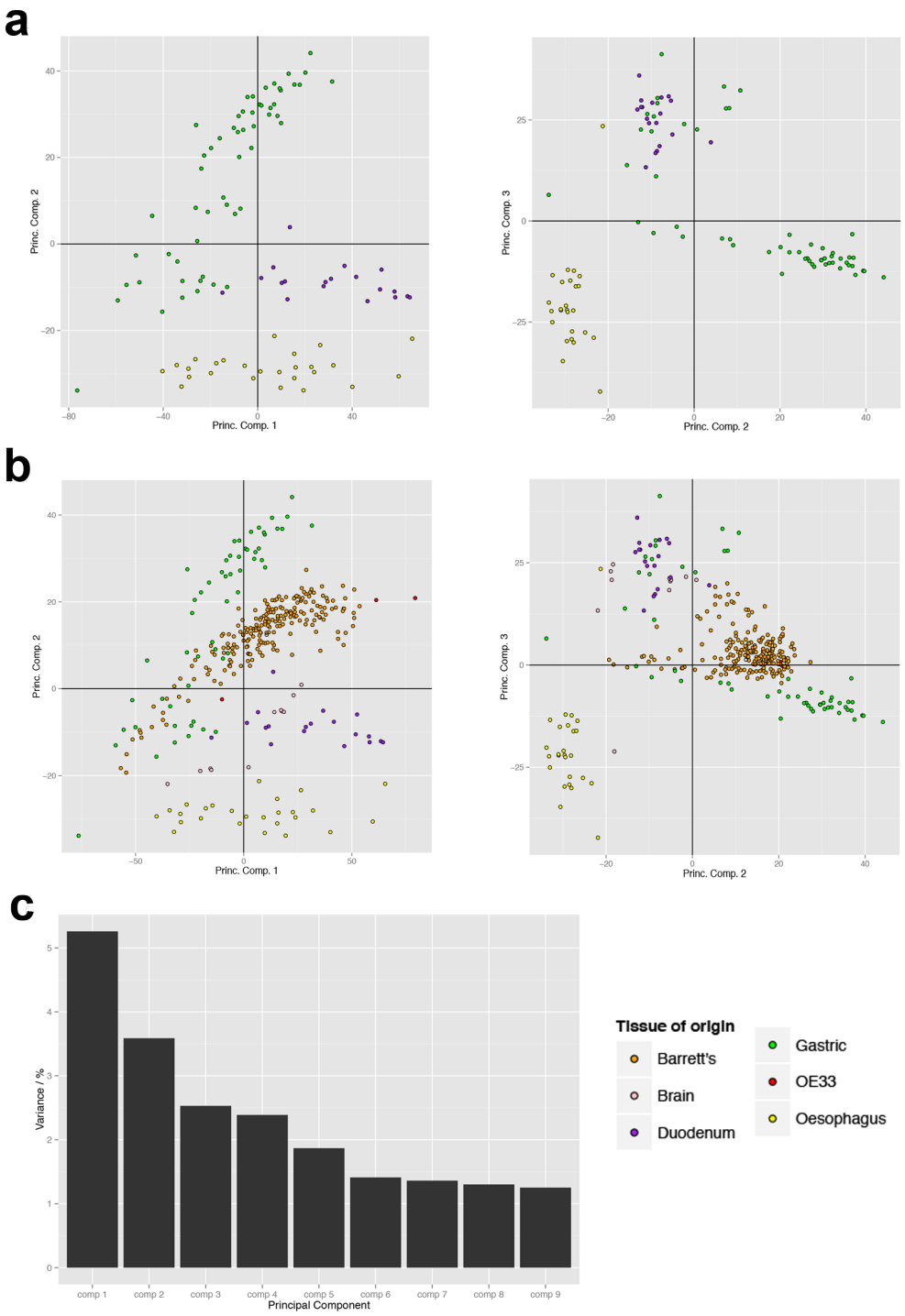


Figure 3.4.8: Tissue type is still the dominant signal in the single cell libraries with duplicates removed. Principal component analysis dimensions 1,2, and 3 of preliminary single cell data with duplicates removed. **a**, Control tissue types informed the vector space rotation and **b** Barrett's samples were treated as supplementary variables and projected onto the vector space in **a**. **c**, The variance explained by each component and the legend for the upper panels.

3.4.6 Duplicates were removed before further analysis of our data

I have demonstrated that a duplicates represent the majority of mapped transcripts in our sequenced single cell libraries. However, on a gene-by-gene basis, the reported gene expression was mostly unchanged for the majority of the expressed genes in the library when duplicates were removed. Reported gene expression (in raw counts), but not transcript length, associated with the extent to which reported gene expression changed when duplicates were removed (Δ). This suggested to me that the duplicate signal in our data was more likely reflective of PCR amplification artifacts rather than biological features. Indeed, it has been noted that high duplication rates are typically PCR amplification artifacts indicative of low cDNA concentration present in the original sample (Bacher & Kendziorski 2016; Grün & van Oudenaarden 2015). Moreover, the overarching biological signal (tissue type) was invariant to duplicate removal. If anything, the signal was observed to be stronger relative to noise when duplicates were removed. Removal of duplicates is a practice described in the literature (Bacher & Kendziorski 2016; Grün & van Oudenaarden 2015) and recommended by Professor Chris Ponting during discussion concerning this project. Therefore, I elected to remove duplicates prior to transcript quantification as part of our single cell pipeline.

3.5 5' or 3' bias was not present in our data.

Although Smartseq2 (Picelli *et al.* 2014) has been reported to minimize 5' and 3' bias in single cell RNA-sequencing experiments, I wanted to ensure that our data did not have any significant read density biases along the length of the transcripts. To that end, I examined the read density along all transcripts. Transcripts were binned into 100 segments of varying length depending on the total length of the transcript. Read counts were computed for each bin on each transcript and the sum total count across all annotated transcripts was reported for each of the 100 bins (Figure 3.5.1). I observed no systematic bias across the body of the transcripts.

Thus, our utilization of Smartseq2 is in agreement with the literature concerning biases across transcripts (Picelli *et al.* 2014).

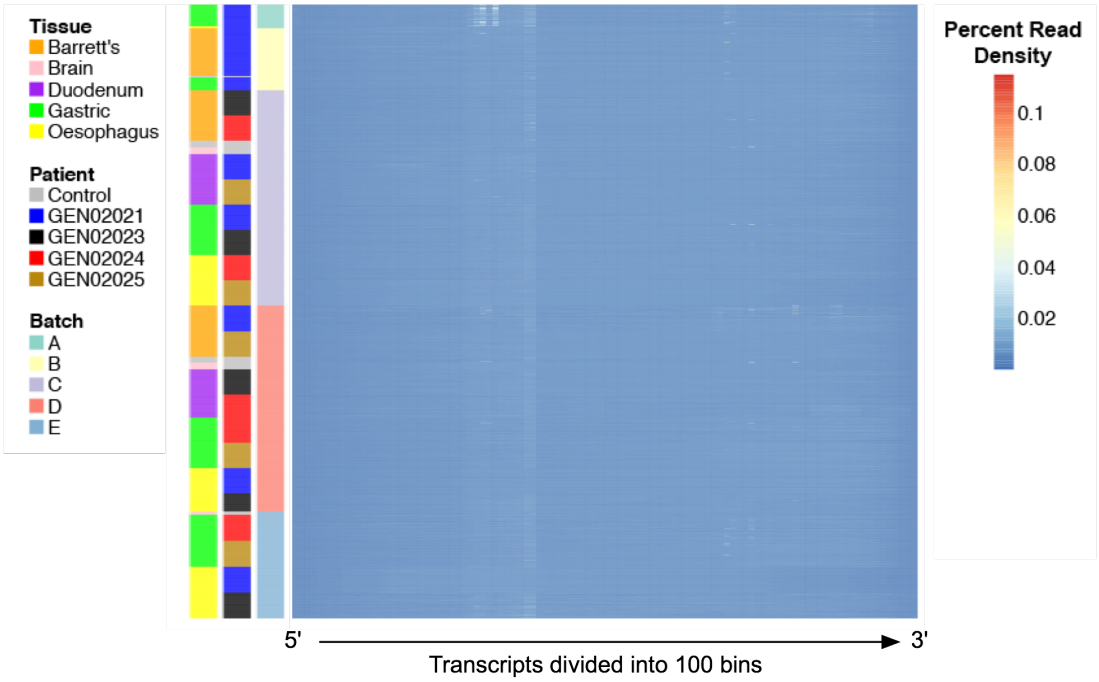


Figure 3.5.1: Heatmap displays the mapped fragment density across all reference genes segmented into 100 bins (columns) for all wells sequenced in batches A-E, which amounts to all cells sequenced from patients that presented with Barrett's esophagus. Each well's tissue, patient, and batch are annotated to the left of the heatmap.

3.6 Mapping metrics and QC of single cell RNA-sequencing libraries

3.6.1 Mapping metrics are not effective indicators of library quality in our data

In order to delineate cells with high quality biological transcriptomes from those that failed, I examined the number of reads mapped to genome. I found that single-cell libraries performed poorly concerning traditional mapping statistics. Specifically, I observed that in 68.91% of the 3051 cells sequenced, or 2104 cells, less than 50% of sequenced fragments mapped to the human genome (Figure 3.6.1). I observed that the the total number of alignments to the genome varied greatly,

ranging from a thousand to over a million alignments (Figure 3.6.2). This is likely a result of the inevitable uneven pooling that occurs in single-cell sequencing experiments. I also found that sequencing depth did not correlate with library size (Figures 3.6.3 and 3.6.4). This suggests that library sizes were limited by cDNA concentration and initial RNA quality rather than sequencing real estate. Furthermore, this observation implies that sequencing more than approximately 750,000 fragments per cell from samples isolated from endoscopic biopsies will not provide additional information concerning the transcriptome.

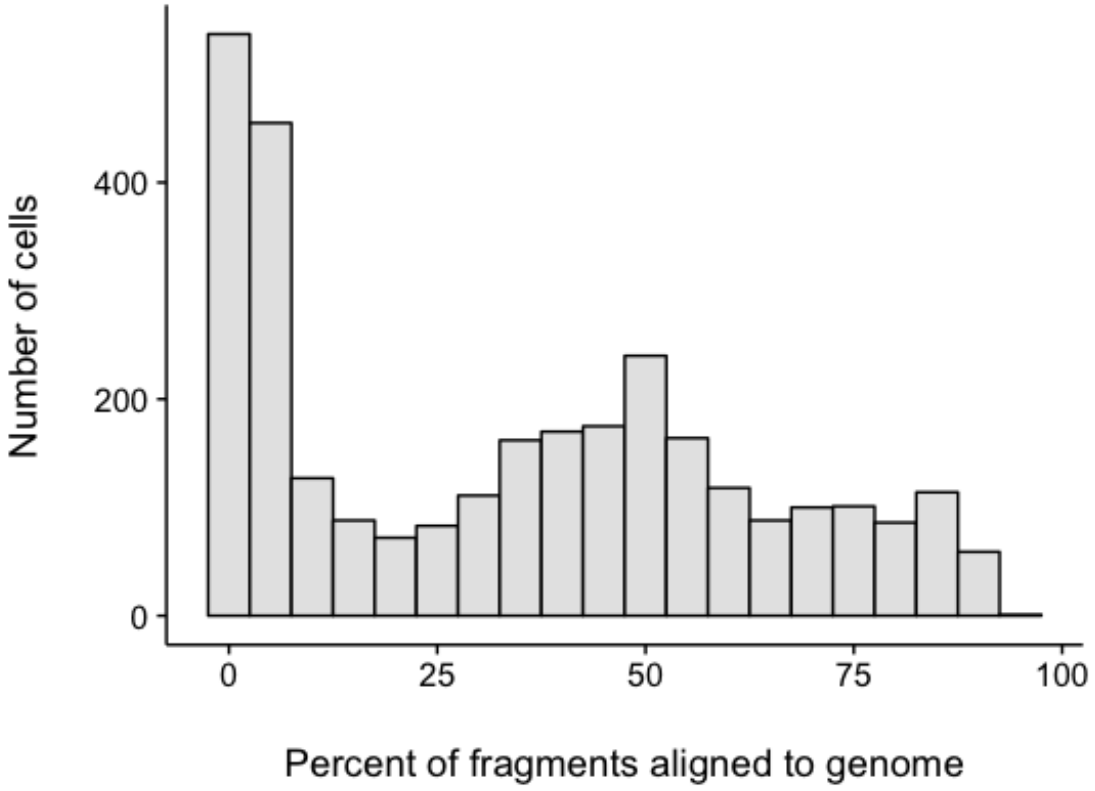


Figure 3.6.1: Histogram illustrates the distribution of the fraction of sequenced fragments aligned to the genome per cell is show. Bars indicate the number of cells (y-axis) observed to have a binned fraction of sequenced fragments aligned to the genome (x-axis).

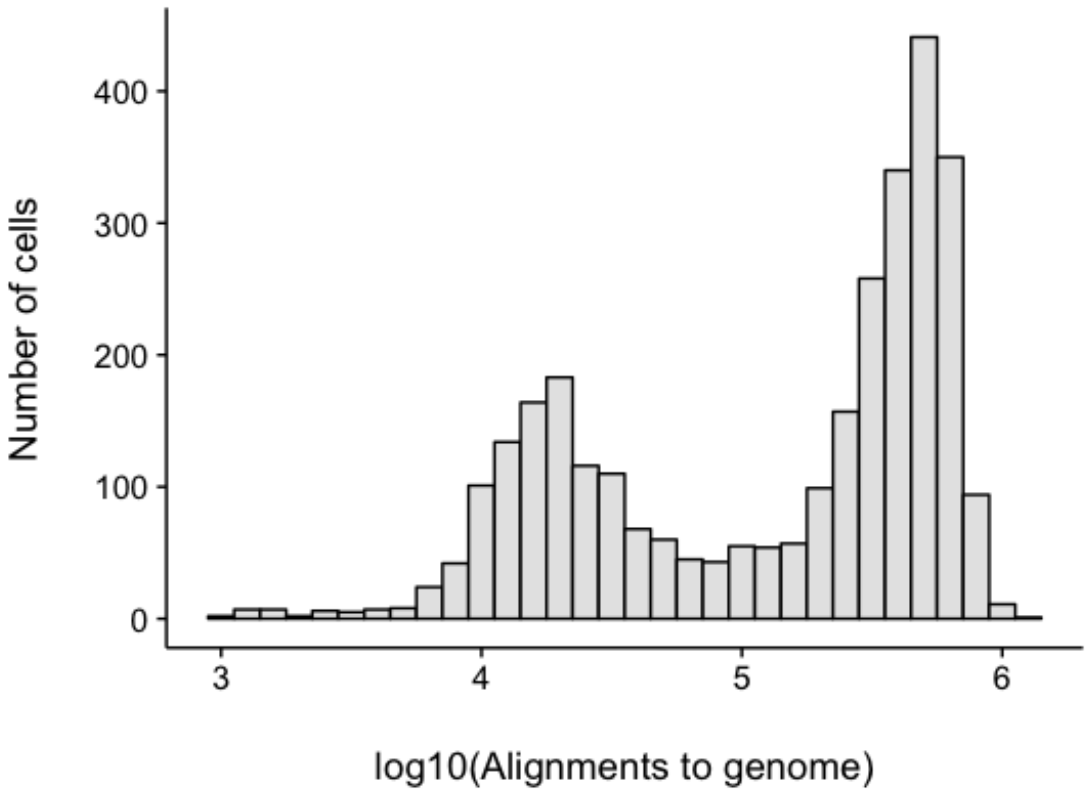


Figure 3.6.2: Histogram illustrates the distribution of the number of genomic alignments per cell is shown. Bars indicate the number of cells (y-axis) observed to have a binned number of genomic alignments (x-axis).

I also observed that blank wells containing only ERCC transcripts and water generally had a lower fraction of fragments map to the genome (ERCC spike-in transcripts are included in the reference genome). These libraries also generally had lower sequencing depth (Figure 3.6.3). In contrast, I observed no relationship between either mapping statistic for other tissue types including the whole tissue brain positive controls (Figure 3.6.3). I did observe a clear bias in sequencing depth across the 5 batches, but the fraction of fragments mapped to the genome showed no such relationship (Figure 3.6.4). Specifically, batches A and B showed particularly high sequencing depth as compared to the other batches. This likely reflects the fact that batches A and B contained only 179 and 221 cells respectively, while the other four batches averaged 662 cells. Thus, batches A and B had more sequencing real estate relative to the libraries in other batches, which were of comparable total cDNA concentration and quality. The clear differences in sequencing depth across

batches reinforces the need to control for batch effects in single cell data. Since the mapping metrics for all samples, including the brain whole tissue positive control samples were generally poor, I elected to examine the behavior of ERCC control molecules and the total number of reads mapped to the transcriptome in order to identify information rich high quality cells for downstream biological analysis.

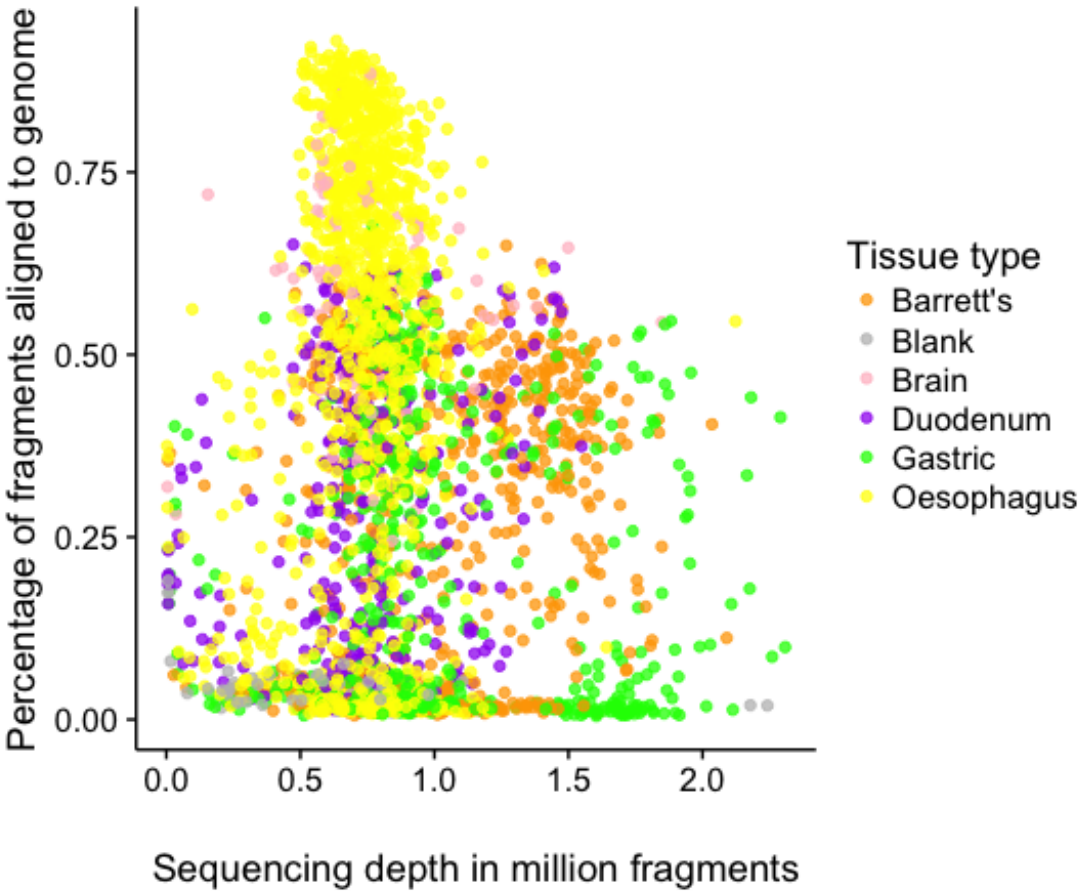


Figure 3.6.3: Scatterplot illustrates the fraction of fragments that aligned to the genome (y-axis) vs the sequencing depth in millions of fragments (x-axis). Points represent a cell library and are colored by tissue or control of origin.

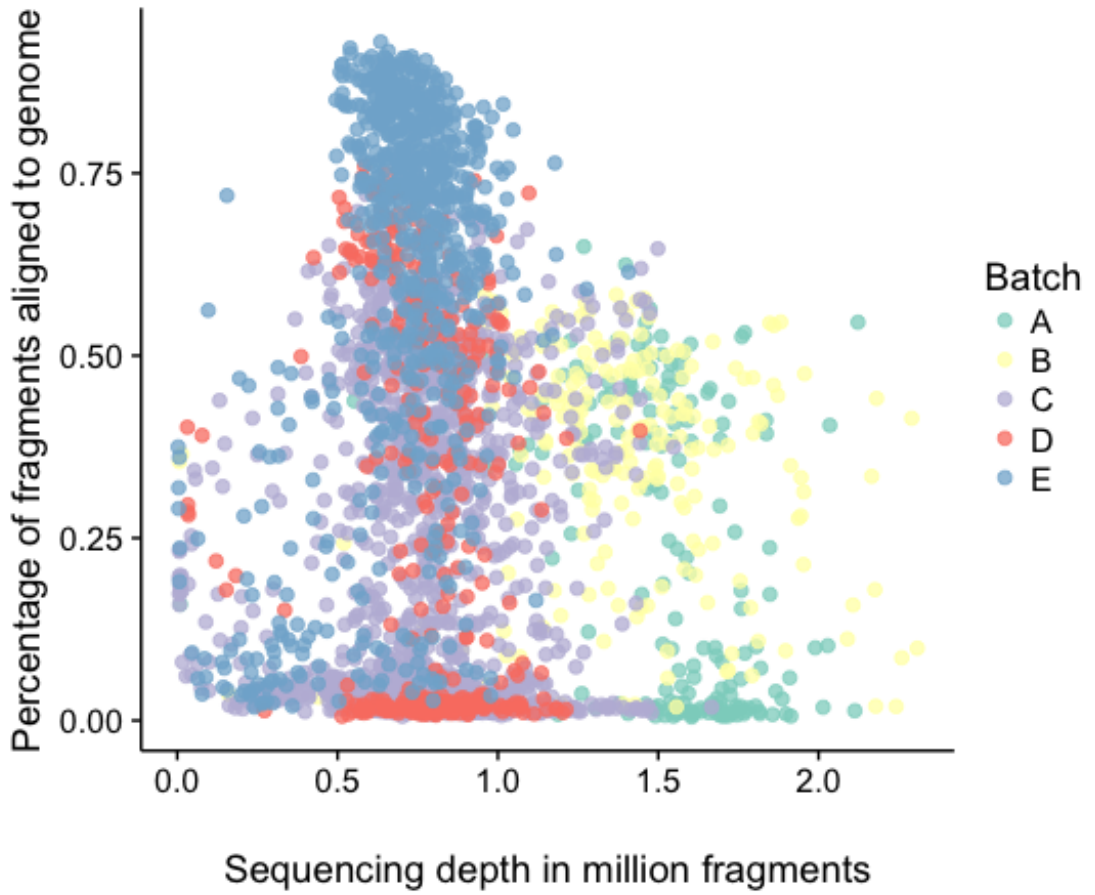


Figure 3.6.4: Scatterplot illustrates the fraction of fragments that aligned to the genome (y-axis) vs the sequencing depth in millions of fragments (x-axis). Points represent a cell library and are colored by batch.

3.6.2 ERCC spike-in concentration in our samples

I reasoned that ERCC spike-ins could be used as a metric for successful sequencing of high quality transcriptomes, because each well contains the same concentration of ERCC spike-ins within experimental variability. I reasoned these transcripts could also provide an initial assessment of the success of the overall experiment (i.e. do ERCC spike-in expression values correlate with expected transcript counts?). I examined the expression distribution in TPM for each ERCC spike-in across all 3051 libraries (Figure 3.6.5). I observed that ERCC spike-in measurements in our libraries were generally low. Only the three ERCC spike-ins with the highest expected transcript number were observed without drop-out events. This is not surprising given the highest expected number of transcripts for a single

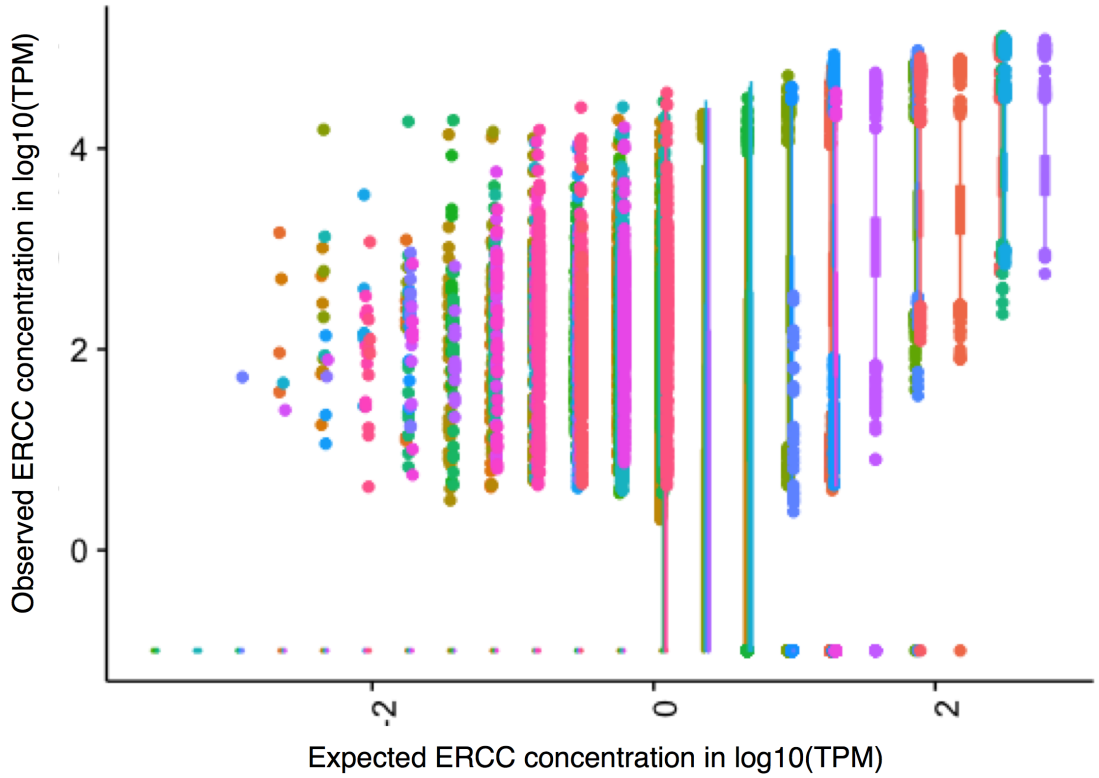


Figure 3.6.5: Boxplots illustrate the distribution of observed TPM of all ERCC spike-ins across cells as function of the expected number of transcripts. Individual ERCC spike-in molecules are separated by color.

ERCC spike-in within a well was only 602. This suggests that future users of this protocol may want to consider adding higher concentrations of ERCC spike-in solution to libraries. However, ERCC spike-in observations with expected transcript numbers greater than 2 did exhibit correlation (Pearson's R: 0.67 and p-value $< 2.2\text{x}10^{-16}$) between observed concentration in TPM and expected transcript number suggesting that library preparation chemistry and sequencing were successful in the experiment. I also examined the number of wells with a drop-out as a function of the expected number of transcripts in the well (Figure 3.6.6). Unsurprisingly, drop-outs were prevalent and occurred for all ERCC spike-in molecules with concentrations below 301 expected transcripts per sample.

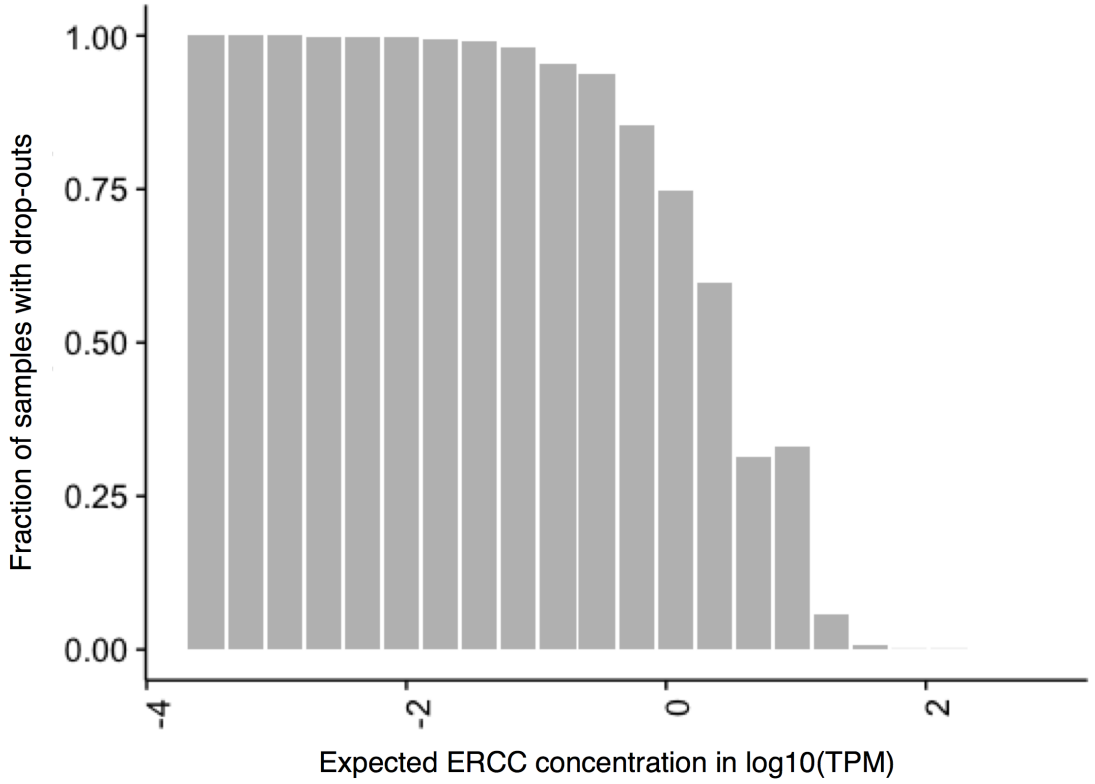


Figure 3.6.6: Drop-outs are prevelant in ERCC spike-ins. Boxplots illustrate the distribution of observed TPM of ERCC spike-ins across cells as function of the expected number of transcripts

I wanted to understand how the percentage of the library occupied by ERCC spike-ins varied with sequencing depth. Therefore, I summed the reported concentration of ERCC spike-ins in TPM and determined the percentage of each library (normalized to one million transcripts) composed of ERCC spike-ins. I examined this value with respect to the sequencing depth of each library in total fragments sequenced (Figure 3.6.7). There was no relationship between sequencing depth and the fraction of the library composed of ERCC spike-ins. This is congruent with the previous observation that the fraction of alignments successfully aligned to the genome was uncorrelated to sequencing depth, and is likely a result of the relatively low concentration and quality of RNA present in samples compared to sequencing real estate. Unsurprisingly, blank water samples appeared to have relatively low sequencing depth and high ERCC spike-in contributions to the observed transcriptome. I also observed that the density of cells was highest near

the extreme values of 0% or 100% regardless of tissue type; intermediate ERCC fraction values were rare across all sequencing depths.

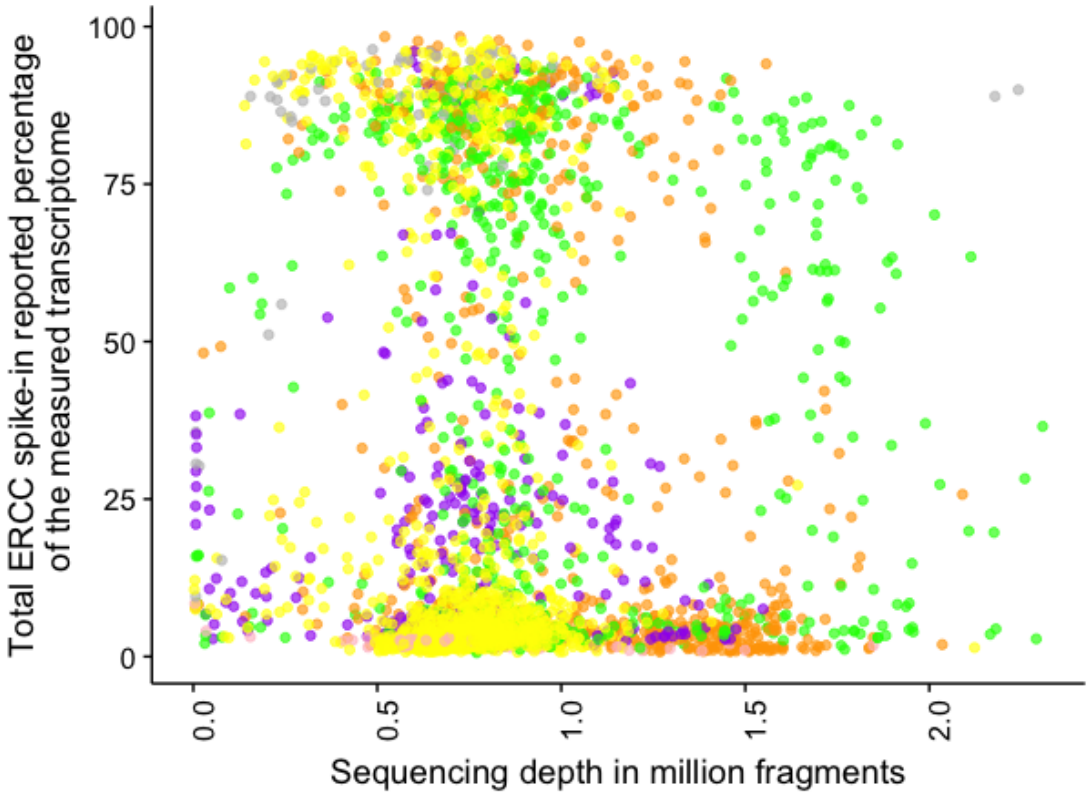


Figure 3.6.7: A scatterplot of the percentage of the measured transcriptome occupied by ERCC spike-ins (y-axis) vs the sequencing depth in total fragments (x-axis). Each point is a sample, and is colored by tissue or control type.

In order to understand the sparse intermediate ERCC spike-in contribution in a tissue and control specific context, I examined the portion of the library occupied by ERCC spike-ins as a function of the tissue or control from which the sample originated. I observed that libraries that originated from endoscopic biopsies had bimodal distributions of ERCC content (Figure 3.6.8). Furthermore, the positive and negative control samples had well separated unimodal distributions of ERCC content. In fact, the positive and negative control sample distributions of fraction ERCC spike-in content were mutually exclusive, where the maximum ERCC content in the positive control, 8.41%, was below the minimum ERCC content in the negative control, 9.52%. Moreover, the medians of the ERCC content

in the positive control, 2.54%, and the negative control, 90.06%, are drastically different. Notably, a small number of blank samples had lower ERCC spike-in contribution to the library, where the two lowest samples had 9.52% and 15.28% respectively. Furthermore these samples were clearly outliers with respect to the rest of blank samples, given that their ERCC content was far below 1.5 times the interquartile range subtracted from the lower quartile, 79.13%. Taken together, these observations suggest ERCC content reflects the extent to which libraries have high quality biological information.

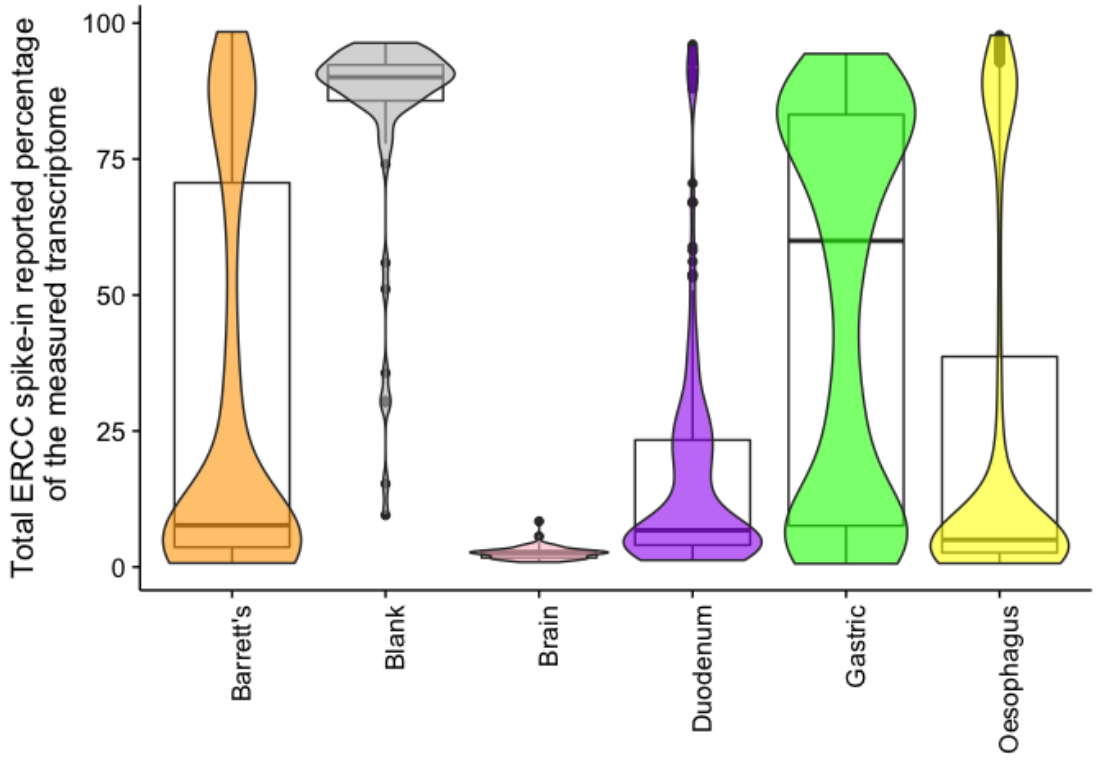


Figure 3.6.8: Boxplots overlaid by violin plots illustrate the distribution of ERCC content in percentage of the observed transcriptome for samples in each tissue or control type. Violin plots are colored by tissue or control of origin and indicate the distribution of samples; the width of violins is normalized across tissue or control types. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

3.7 Mitochondrial and rRNA expression in our samples

In order to understand the quality of our single cell RNA-sequencing samples I investigated the rRNA and mitochondrial gene expression of samples. Furthermore, I compared the single cell samples' rRNA and mitochondrial expression to those found in whole tissue samples from the four patients that presented with Barrett's esophagus. As with ERCC content, I examined the fraction of the transcriptome (as reported in TPM) composed of rRNA and mitochondrial transcripts.

3.7.1 Nuclear rRNA gene expression is low in our samples

I examined the rRNA content of single cell transcriptomes and dilute whole tissue brain (positive) and water blank (negative) controls. By rRNA content I refer to the all genes encoding ribosomal RNA that are transcribed within the nucleus of the cell. The SMARTseq2 protocol enriches for polyadenylated transcripts; therefore, I expected rRNA content to be low in single cell libraries. Indeed, I observed that rRNA expression was extremely low across all samples in our single cell RNA-sequencing experiment (Figure 3.7.1). No sample had greater than 0.4% rRNA content and with the exception of a few outlier samples, rRNA content was below 0.02%.

I also expected rRNA content to be low in whole tissue samples, because these samples were subject to a ribo-depletion protocol, which should degrade any rRNA present within the cell. I observed that ribo-depletion of rRNA transcribed in the nucleus was successful, where rRNA content was generally below 0.0025% of each transcriptome (Figure 3.7.2). For two patients, GEN02021 and GEN02023, rRNA content in samples from the duodenum was noticeably higher than all other samples, but rRNA content was still below 0.0065% of the transcriptome. Collectively these observations suggest that single cell libraries' rRNA content were consistent with that observed in whole tissue libraries. Furthermore, because I observed extremely low rRNA content present across all samples, whole tissue and single cell, I decided that rRNA content was not an issue for further analysis.

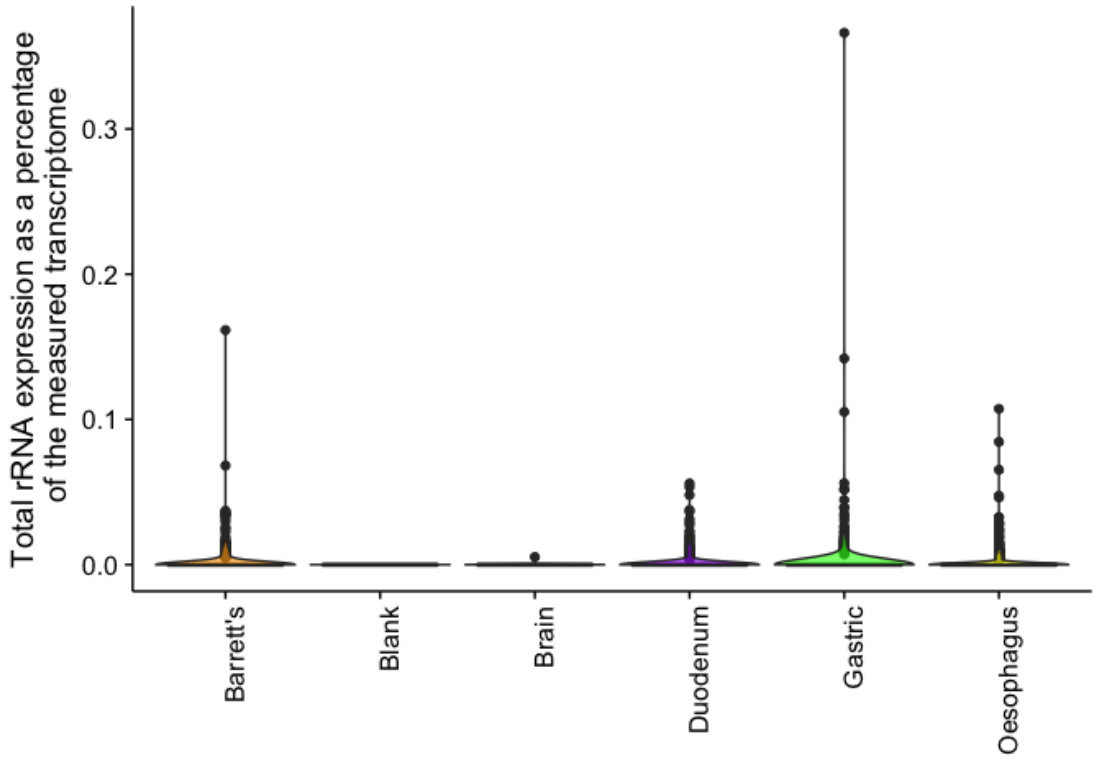


Figure 3.7.1: Boxplots overlaid by violin plots illustrate the distribution of rRNA content as a percentage of the observed transcriptome for samples in each tissue or control type. Violin plots are colored by tissue or control of origin and indicate the distribution of samples; the width of violins is normalized across tissue or control types. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

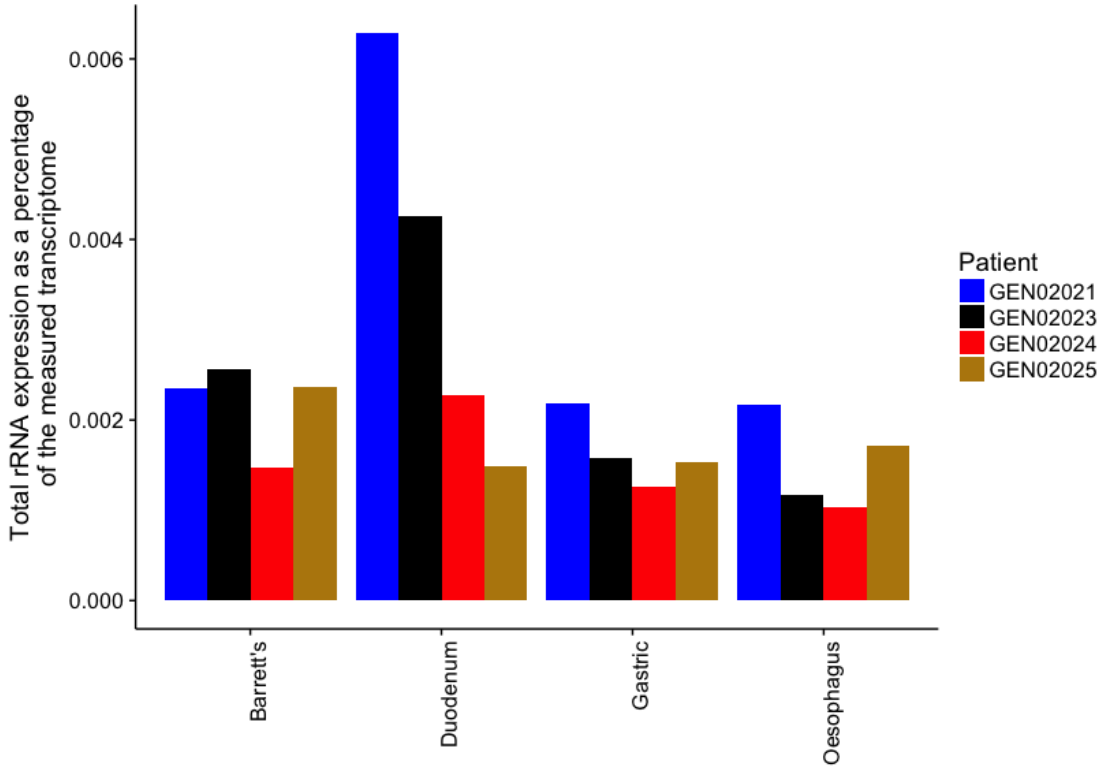


Figure 3.7.2: Bars indicate the fraction of the transcriptome occupied by rRNA transcribed from the nucleus in 16 whole tissue samples from patients presenting with Barrett's esophagus. Tissue source is indicated on the x-axis and patients are indicated by bar color.

3.7.2 Mitochondrial gene expression is high in our libraries

In addition to rRNA content, I examined the mitochondrial expression content in single cell transcriptomes and dilute whole tissue brain (positive) and water blank (negative) controls. By mitochondrial expression, I refer to all RNA molecules that are transcribed in the mitochondria and not the nucleus. There are 37 such genes: 2 rRNA, 13 protein coding, and 22 tRNA genes. Overall, mitochondrial gene expression composed a large proportion of the libraries, where the average mitochondrial RNA content as a fraction of the transcriptome was 51.54% for samples from endoscopic biopsies, and 30.93% for control samples (Figure 3.7.3). I observed that for the majority of cells from endoscopic biopsies, mitochondrial expression as a percentage of the transcriptome was on par with the dilute whole tissue brain positive control. Interestingly, the distribution of mitochondrial content across cells clearly varied by tissue type; cells from the tissue with most specialized epithelia (the duodenum) were more likely to have high mitochondrial content, and cells from the least specialized (the proximal normal esophagus) were less likely to have high mitochondrial content. The distributions of cell mitochondrial content in the Barrett's lesion and gastric tissue existed on a continuum between the two extremes, where cells from gastric tissue were more likely to have high mitochondrial content.

Unsurprisingly, blank samples contained the least mitochondrial content; the median mitochondrial content was 3.55%. Intriguingly, I observed that a handful of blank samples with more than 10% mitochondrial content. Having observed a similar number of outlier samples with surprisingly low ERCC spike-in content, I directly compared the ERCC spike-in and mitochondrial content of blank samples (Figure 3.7.4). The ERCC spike-in and the mitochondrial content in blank samples were highly correlated (Pearson's R: -0.929 and p-value $< 2.2\text{x}10^{-16}$). Furthermore, the outlier samples with low ERCC spike-in content were the same outlier samples with aberrantly high mitochondrial content. I suspected that these samples were contaminated with mitochondria, and thus should not be used for further analysis. Positive and negative control samples were included in the experiment in order to delineate high quality samples from failed ones, and these blank samples could

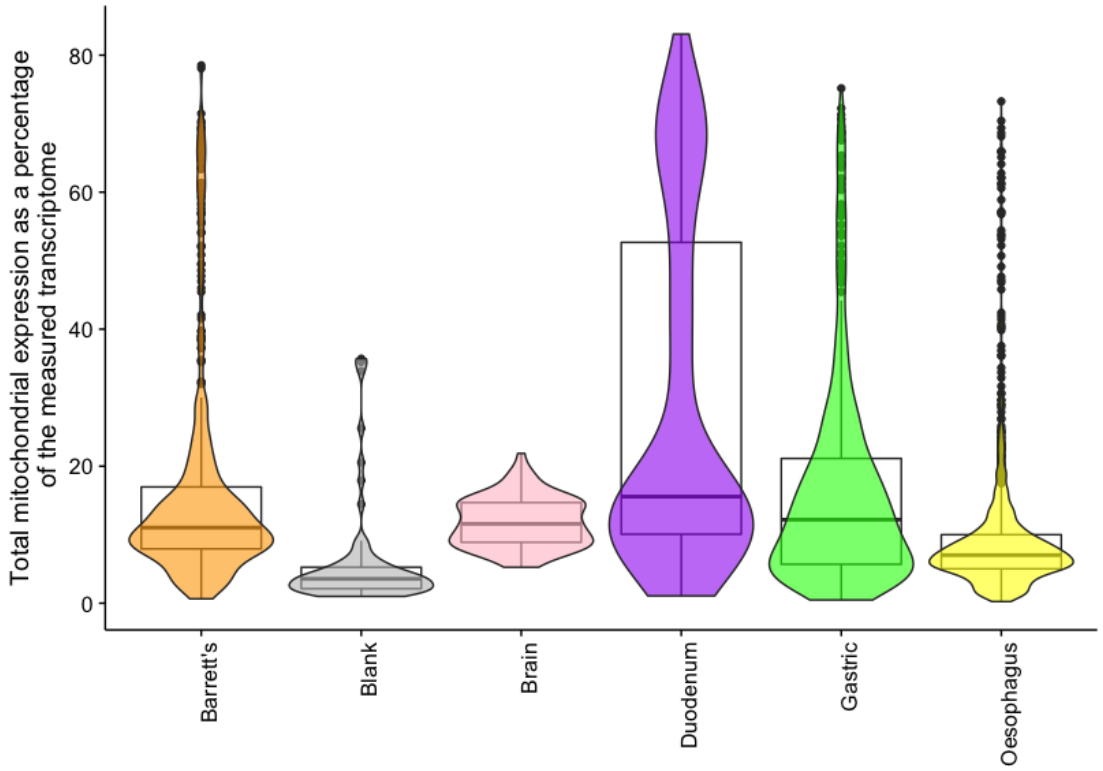


Figure 3.7.3: Boxplots overlaid by violin plots illustrate the distribution of mitochondrial expression as a percentage of the observed transcriptome for samples in each tissue or control type. Violin plots are colored by tissue or control of origin and indicate the distribution of samples; the width of violins is normalized across tissue or control types. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

confound this purpose; therefore, the 7 blank samples with greater than 10% mitochondrial content were discarded for all further analysis. This left 64 blank negative control samples for downstream analysis.

I was concerned that the exceptionally high mitochondrial content (greater than 50%) observed in many cells from the duodenum and gastric tissue and observed in a smaller number of cells form the Barrett's lesion was an indication that cells were stressed, or even apoptotic, as a result of the single cell isolation protocol. To explore this further, I examined the mitochondrial content in whole tissue RNA-sequencing samples extracted from tissue proximal to the tissue used for single cell sorting in the same endoscopic biopsies in patients presenting with Barrett's esophagus

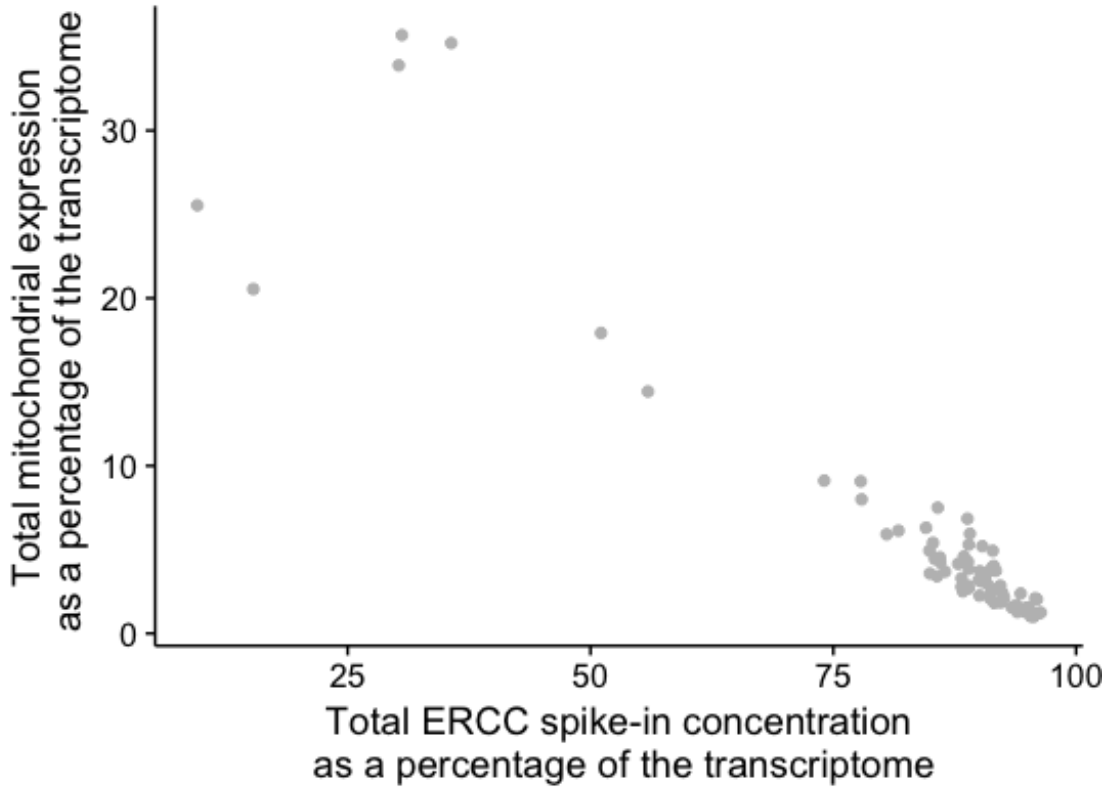


Figure 3.7.4: Scatterplot displays the observed mitochondrial content (y-axis) vs. the observed ERCC content (x-axis) in blank sample libraries.

(Figure 3.7.5). The tissue-specific pattern of mitochondrial content observed in the single cell library samples was qualitatively preserved in the whole tissue sample libraries (Figures 3.7.3 and 3.7.5). Thus, I concluded that mitochondrial content was, in fact, a biologically meaningful feature of the single-cell RNA-sequencing libraries. As a result I decided against any filtering of single-cell samples on the basis of mitochondrial read content.

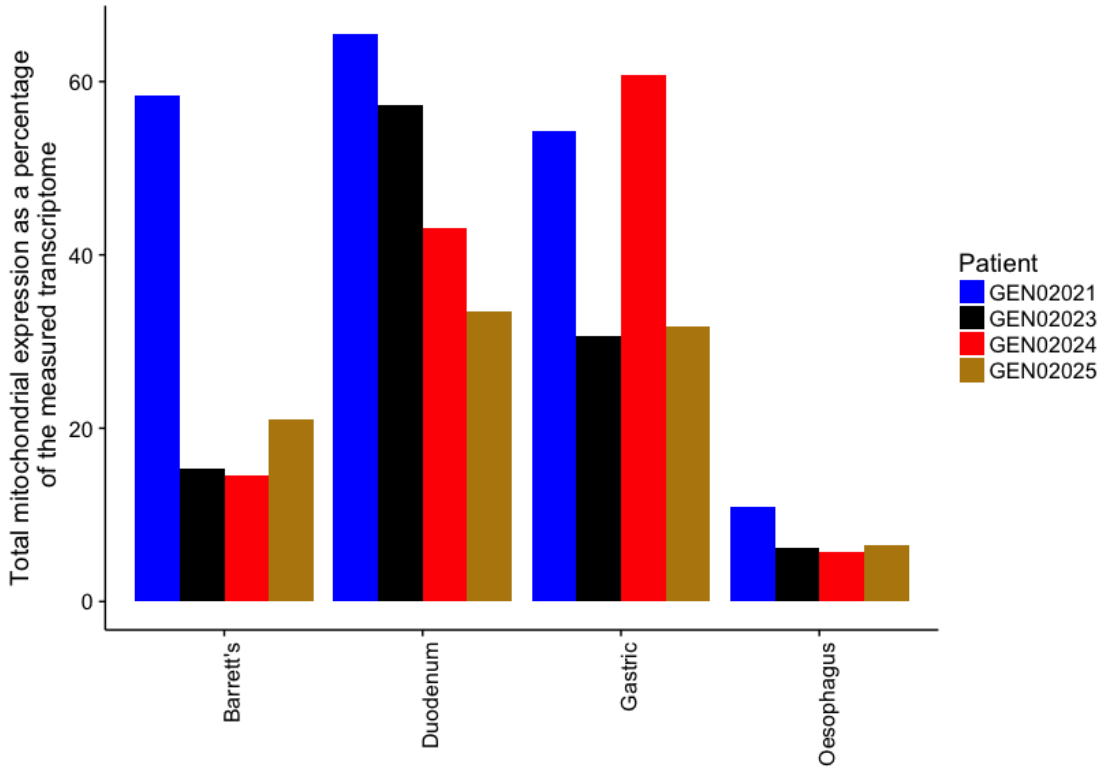


Figure 3.7.5: Bars indicate the fraction of the transcriptome occupied by mitochondrial gene expression in 16 whole tissue samples from patients presenting with Barrett's esophagus. Tissue source is indicated on the x-axis and patients are indicated by bar color.

3.8 High quality libraries were obtained from endoscopic biopsies.

3.8.1 Transcriptome fragment count determines high quality libraries

After obtaining a general understanding of the mapping and transcriptomic features present in our single cell libraries, I wanted to examine the total number of fragments mapped to the transcriptome. I reasoned that high quality libraries would have higher numbers of successful alignments to the transcriptome than failed libraries, and I expected positive and negative control samples to inform our threshold. I observed that the distribution of fragments mapped to the transcriptome was clearly bimodal for samples obtained from endoscopic biopsies and unimodal for

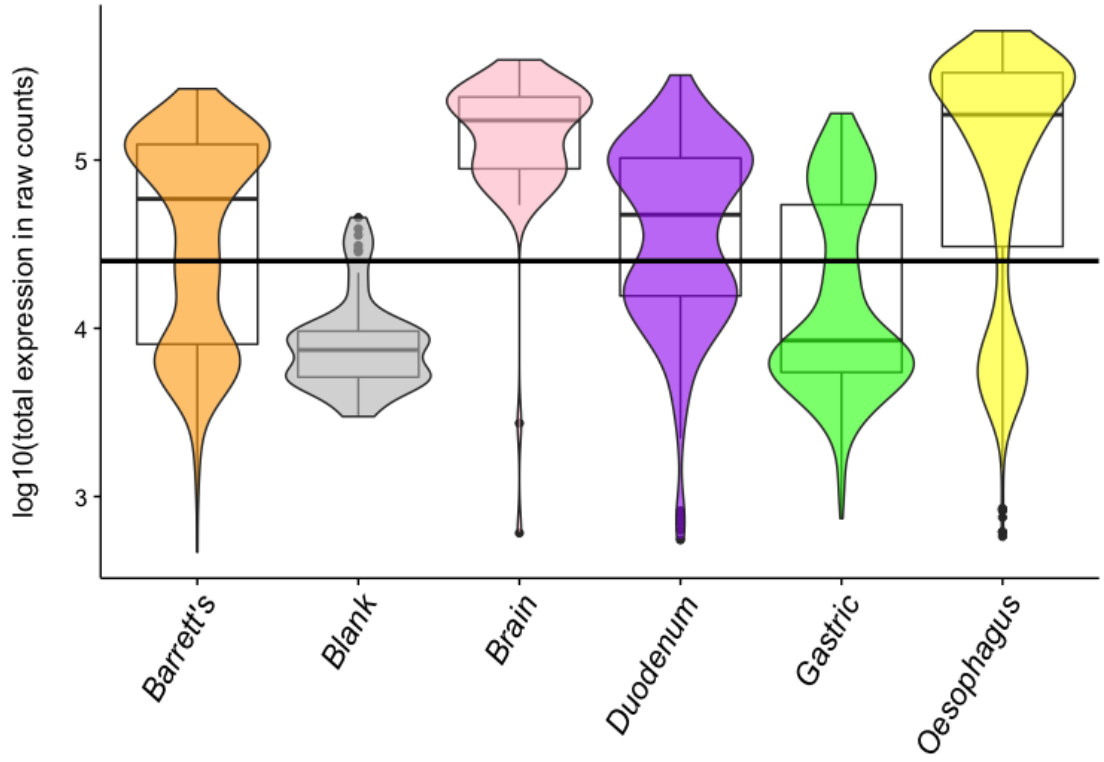


Figure 3.8.1: Boxplots overlaid by violin plots illustrate the distribution of total fragments mapped to the transcriptome for samples in each tissue or control type. Violin plots are colored by tissue or control of origin and indicate the distribution of samples; the width of violins is normalized across tissue or control types. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The bold bar spanning tissue types indicates the threshold defining high quality and failed libraries

positive and negative control samples (Figure 3.8.1). Moreover, the positive control samples had an order of magnitude more fragments aligned to the transcriptome than the blank samples. Excluding outlier samples (plus or minus 1.5 times the interquartile range for the upper and lower quartiles, respectively), the minimum observed fragments mapped to the transcriptome was 53,703 for brain positive control samples and the maximum observed fragments mapped to the transcriptome was 21,380 for blank negative control samples.

Furthermore, I suspected that the number of reads mapped to the transcriptome was a reasonable proxy for library quality, given the bimodal distributions in samples

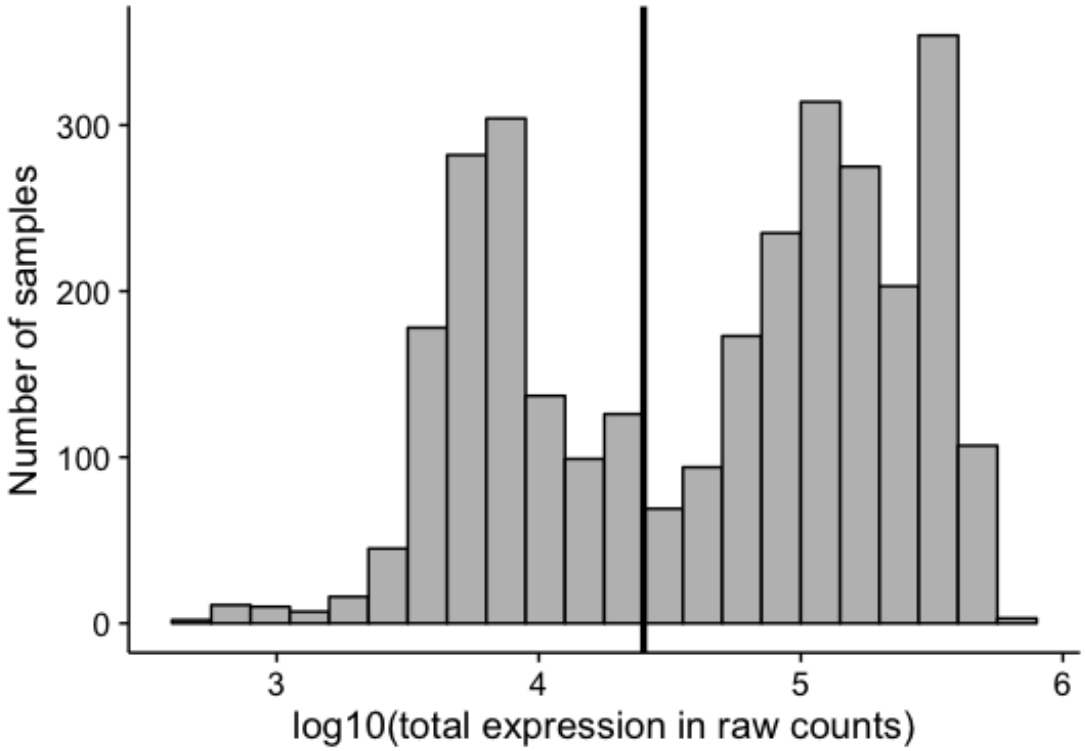


Figure 3.8.2: Histogram illustrating the distribution of library sizes as measured by fragments mapped to the transcriptome.

from endoscopic biopsies and the clear separation between the unimodal positive and negative control distributions. To that end, I examined the distribution of total fragments mapped to the transcriptome across all sample types and visually determined a putative threshold to delineate high quality and failed single cell libraries (Figure 3.8.2). All samples from endoscopic biopsies and positive control samples with greater than 25,118 fragments mapped to the transcriptome were considered high quality and all samples below this threshold were considered failures. Outlier blank samples above the threshold were not considered as high quality transcriptomes and were excluded from further analysis. Notably, the selected threshold clearly splits the bimodal distributions of total fragments mapped to the transcriptome for each tissue and seperates the unimodal brain and blank distributions (Figure 3.8.1).

3.8.2 Transcriptome fragment count and library complexity

In order to verify that the threshold set on total fragments mapped to the transcriptome was a reasonable threshold for determining high quality single-cell transcriptomes for further analysis, I examined the library complexity in samples included as high quality libraries or excluded under the threshold. In order to describe library complexity in included and excluded samples, I took each cell and ordered genes from the highest to the lowest expressing gene and computed the cumulative sum across the 43,197 genes detected across all samples in the experiment. Plotting the genes in decreasing order and their respective cumulative fraction of the transcriptome provides a clear visual description of library complexity; samples with high library complexity would require more genes before asymptotically approaching 100% transcriptome and samples with low library complexity would reach the 100% asymptote with only a handful of genes. I examined the library complexity in this way in 20 randomly selected samples from each tissue, where 10 samples were those excluded as quality control failed samples and 10 samples were those included as quality control passing samples (Figure 3.8.4). With few exceptions excluded cells have extremely low library complexity, while included cells generally have high library complexity in every sample type except the blank negative control, which by definition only consists of excluded cells and have universally low library complexity.

I wanted to understand how library complexity looked on average across samples that were excluded or included as function of the quality control threshold in each tissue type. To that end, I averaged the cumulative fraction of the transcriptome for each indexed gene within tissue types for samples excluded or included as a result of the quality control threshold (Figure 3.8.3). I found that, on average, excluded libraries had extremely low library complexity, while quality control passing samples had much higher library complexity. This was true across all sample types. In order to summarize library complexity for each sample, I identified the minimum number of genes required to cover 75% of all expression within a sample for each sample (Figure 3.8.5). The bulk of samples considered to have passed quality control

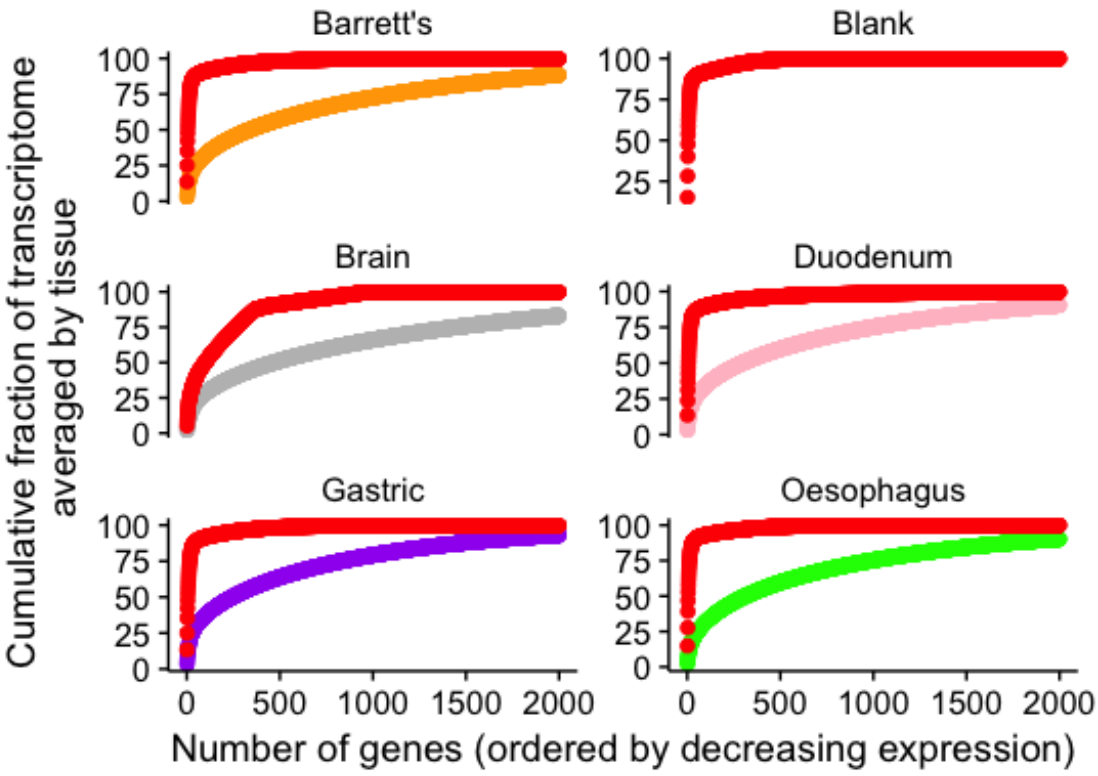


Figure 3.8.3: Scatterplots indicates library complexity by displaying the average fraction of total expression covered (y-axis) by the cumulative expression of some number of genes ordered by decreasing expression (x-axis) for excluded failed samples (red) and quality control passing samples (respective sample color) as determined by transcript number threshold for each sample type (separate panels). All blank negative control samples were excluded.

for each tissue type had much higher numbers of genes required to cover three quarters of their expressed transcriptome as compared to those that were excluded by the threshold. Collectively, these three perspectives on library complexity suggest that the prescribed expression threshold separates complex libraries from those containing little to know information. Finally, I assessed the number of samples with some total number of genes expressed at three thresholds of coverage: 1x, 3x, and 10x (Figure 3.8.6). The number of genes expressed in quality control passing samples was universally higher than the number of genes expressed in the failed samples. Given these observations, I concluded that the threshold is reasonable from the perspective of library complexity.

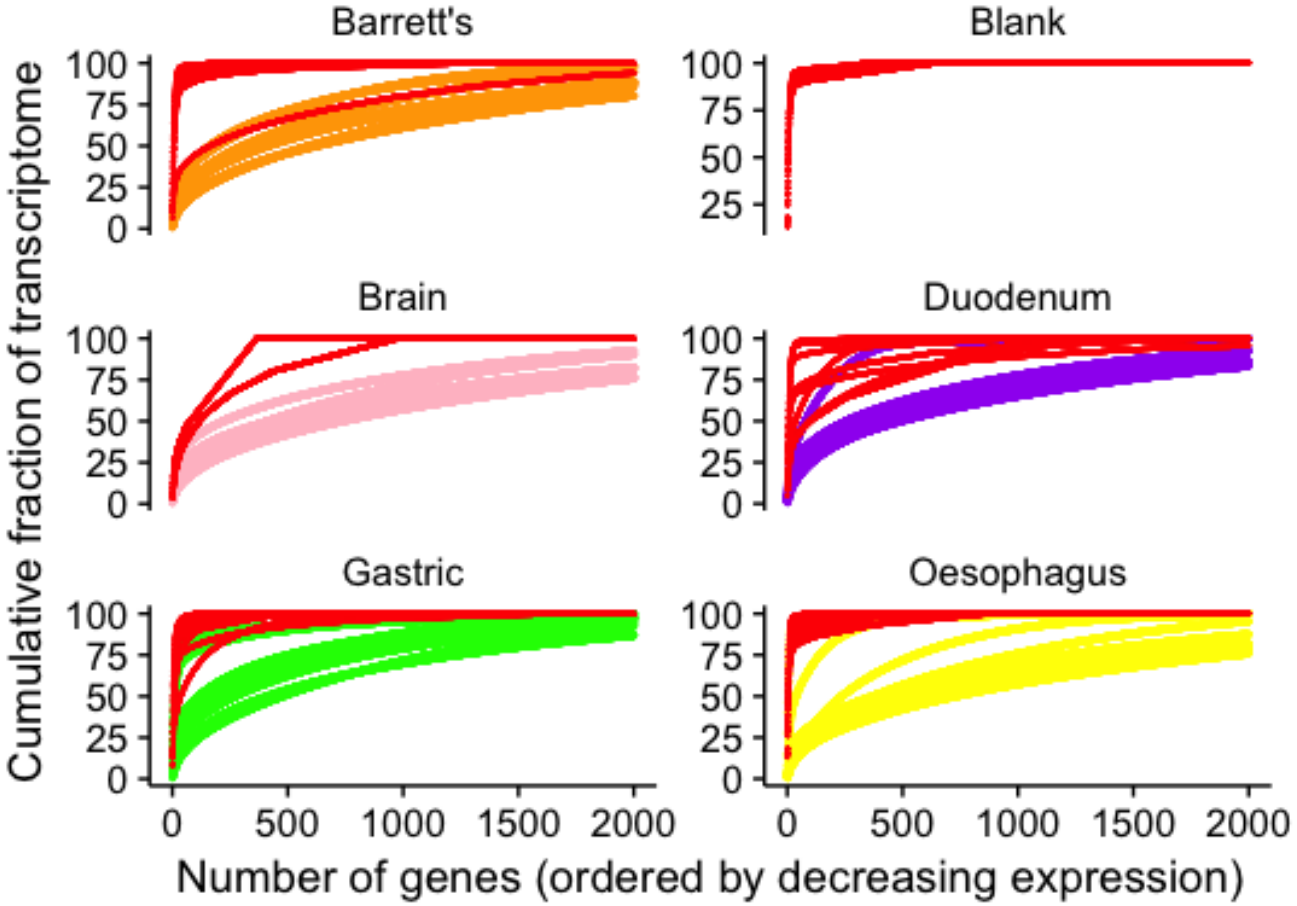


Figure 3.8.4: Scatterplots indicates library complexity by displaying the fraction of total expression covered (y-axis) by the cumulative expression of the 2000 highest expressing genes ordered by decreasing expression (x-axis) for 10 randomly selected excluded samples (red) and 10 randomly selected quality control passing samples (respective sample color) for each sample type (separate panels). All blank negative control samples were excluded.

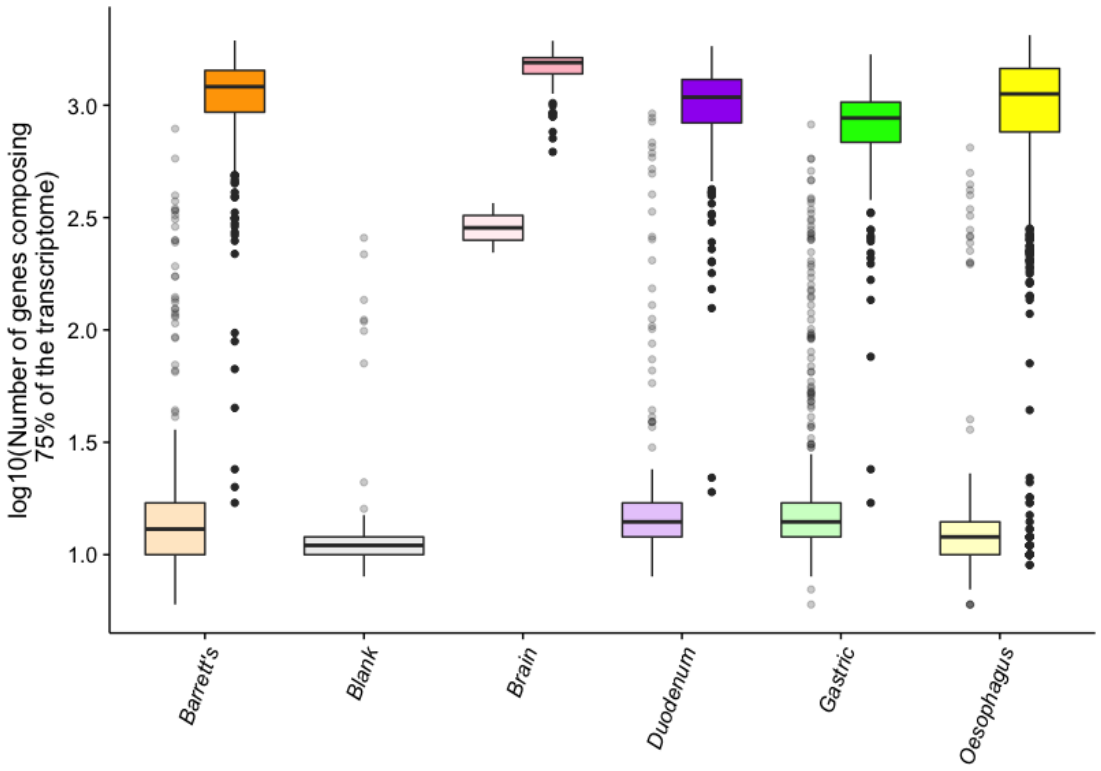


Figure 3.8.5: Boxplots display the distribution of library complexity across cells for each sample type (x-axis), where the distributions for excluded cells (translucent) and included cells (opaque) are shown separately. Library complexity is represented here by the minimum number of genes in the library required to cumulatively sum to 75% of a sample's total expression. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

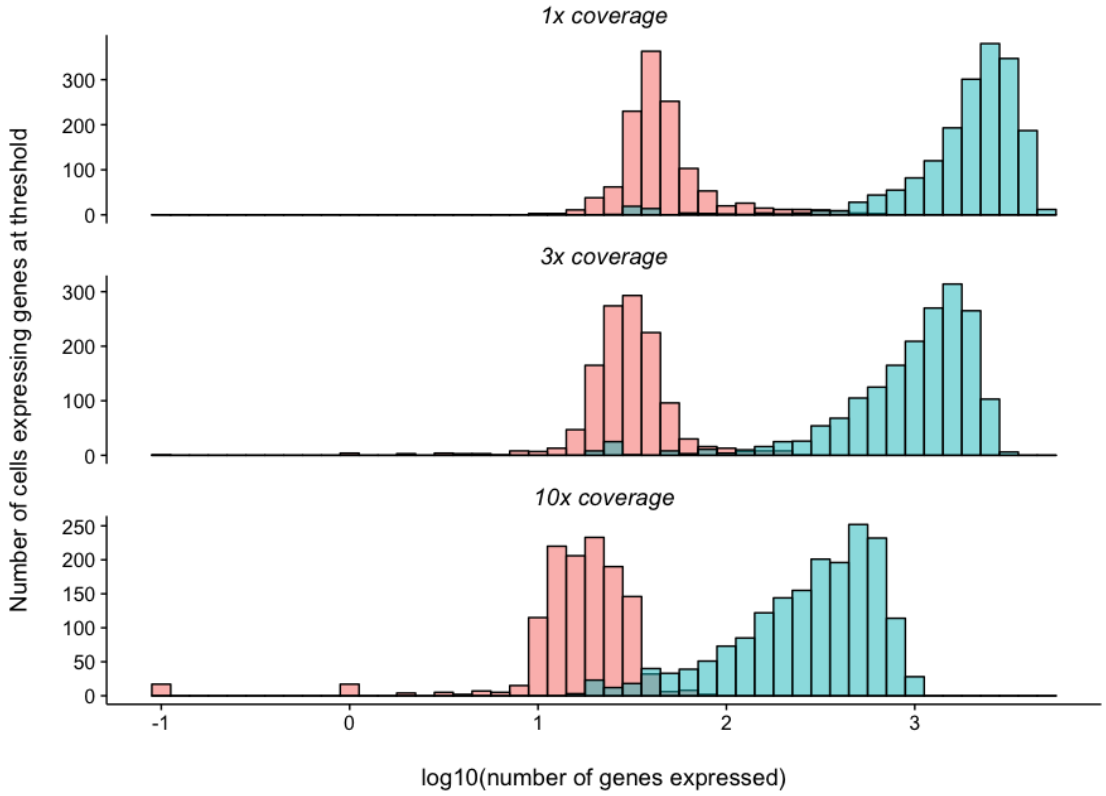


Figure 3.8.6: Histograms display the number of cells (y-axis) expressing a given number of genes (x-axis) at three different depth thresholds for expression (panels): 1x coverage, 3x coverage, and 10x coverage. The red distribution consists of samples excluded as failed samples and the blue distribution consists of samples considered to have passed quality control.

3.8.3 ERCC content is higher in low transcriptome fragment libraries

I previously observed in all 3051 samples that for single cells from endoscopic biopsies ($n = 2849$ cells) the distribution of the fraction of the cells' library taken up by ERCC spike-ins was bimodal; I also noted that the negative control samples ($n = 77$) had exceptionally high proportions of their library taken up by ERCC spike-ins and positive control samples ($n = 85$) had low proportions of their library taken up by ERCC spike-ins (Figure 3.6.8). I reasoned that an effective threshold on the number fragments mapped to the transcriptome would primarily exclude libraries dominated by ERCC spike-ins and retain samples with more endogenous fragments as a fraction of their library. Therefore I compared, within sample types, the the ERCC content in samples that were excluded as failed to the ERCC content in samples included as high quality samples determined by the 25,118 fragment threshold (Figure 3.8.7). I found that the bimodal distributions of ERCC content originally observed across all samples were well split between samples above and below the threshold. Specifically, the libraries excluded as failed samples were overwhelmingly the ERCC-dominated libraries, while libraries included in the group of quality control filtered samples had much higher fractions of endogenous gene expression. I took this as further evidence that the previously discussed threshold is a reasonable one for determining cells with high quality transcriptomes.

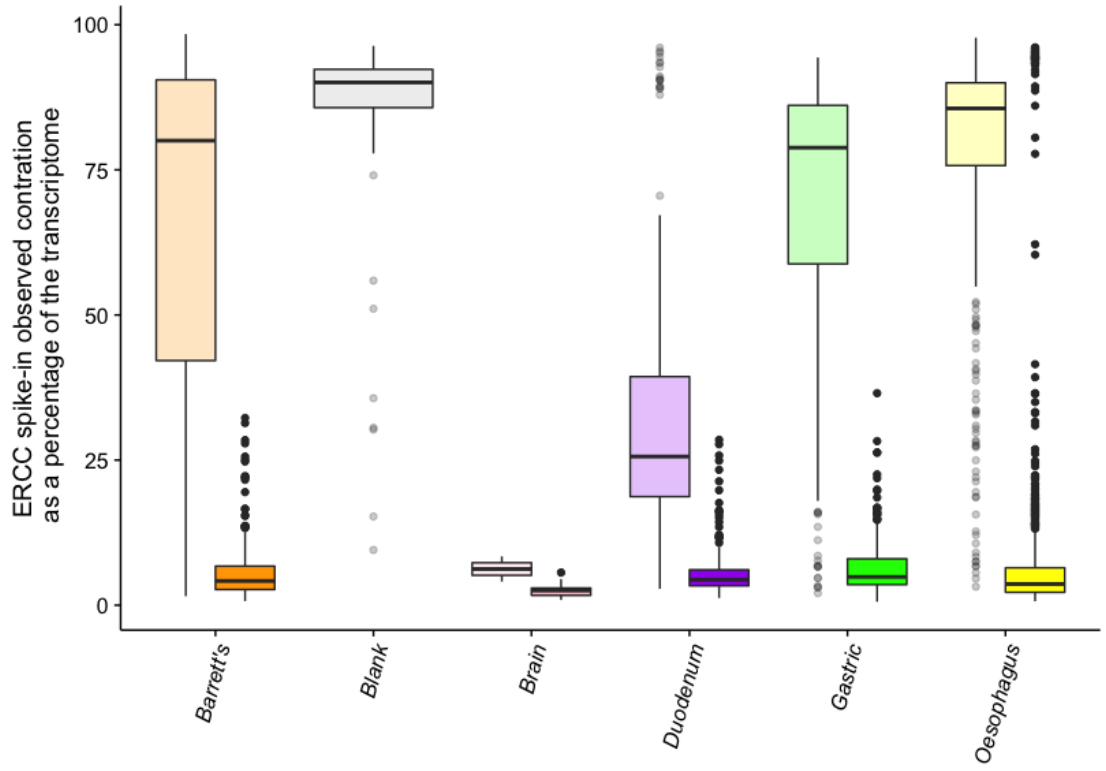


Figure 3.8.7: Boxplots display the distribution of ERCC content (fraction of the library taken up by ERCC spike-ins) across cells for each sample type (x-axis), where the distributions for excluded cells (translucent) and included cells (opaque) are shown separately. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

3.8.4 Feature selection for clustering analysis

Highly expressed invariant genes do not add useful information in terms of defining cells and these genes could result in clustering algorithms being more sensitive to batch effects. Therefore, I elected to remove any genes where both its coefficient of variance was in the bottom decile and its mean expression in the top 5^{th}percentile of quality filtered cells (Figure 3.8.8). These thresholds were determined based on advice in the SC3 algorithm readme; however, the version of the readme suggesting these thresholds is no longer accessible. Thus, although these thresholds exclude the bulk of highly expressed genes with low variation, in hindsight more work could be done to determine the optimal thresholds for feature selection.

These genes were excluded from any clustering analysis in the observed counts and each simulated technical replicate was subject to identical feature selection prior to clustering. Many of the filtered genes consisted of mitochondrial and ERCC spike-in transcripts (Figure 3.8.8). These genes were only filtered for the purpose of clustering, and were still considered when interpreting the biology of single cell clusters.

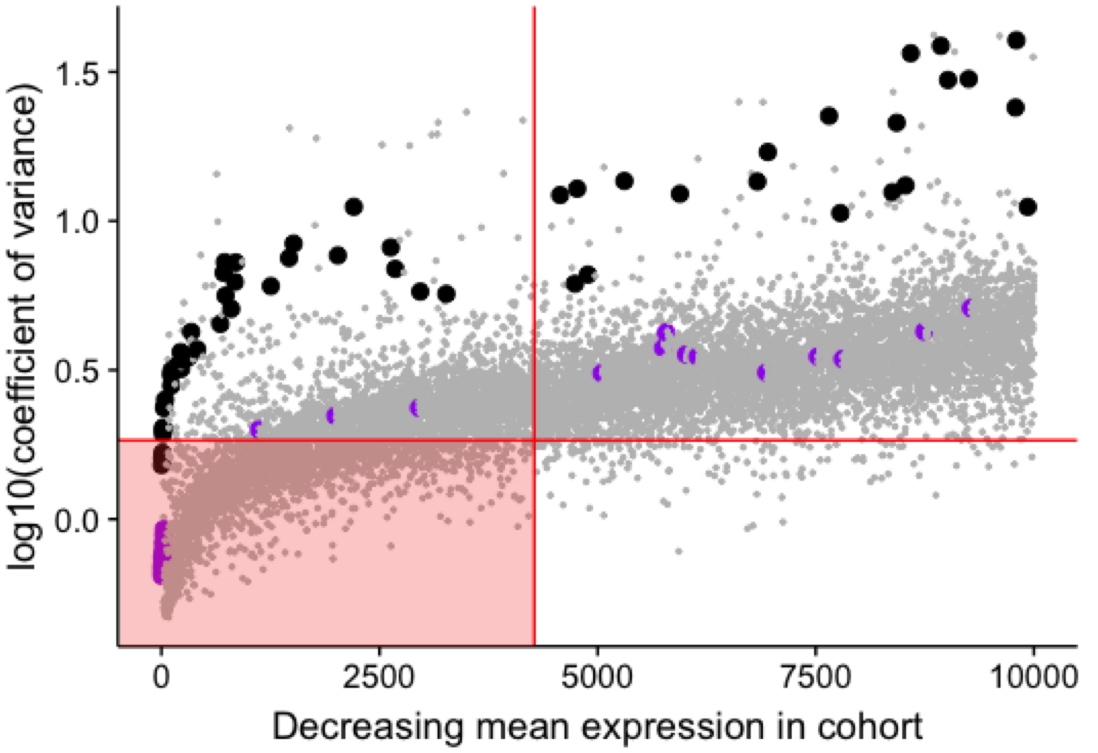


Figure 3.8.8: Scatterplot displays the coefficients of variance for each gene (y-axis), where genes are ordered by mean expression in TPM from highest to lowest (x-axis). Transparent red box denotes the genes excluded from clustering analysis. Genes are colored to indicate transcript type: nuclear transcribed endogenous (grey), mitochondrial transcribed endogenous (purple), ERCC spike-in (black). The horizontal and vertical red lines indicate the thresholds for genes' coefficients of variance (horizontal) and mean expression (vertical).

3.8.5 1821 samples have high quality transcriptomes

I concluded that the threshold of 25,118 fragments mapped to the transcriptome was optimal for seperating the high quality sample libraries from the failed ones. After applying this threshold, I found that 1821 of the 3051 (59.68%) samples sequenced consisted of high quality libraries (Figure 3.8.9). Specifically, 83 of the 85

(97.65%) diluted whole brain tissue samples were considered high quality samples and all of the blank negative controls were explicitly excluded from further analysis. Only 6 of 71 (8.45%) blank negative control samples were above the selected threshold. For samples from endoscopic biopsies from the patients presenting with Barrett's esophagus, 248 of 526 (47.15%) proximal normal esophageal cells, 221 of 678 (32.60%) gastric cells, 207 of 352 (58.81%) duodenal cells, and 371 of 620 (59.84%) cells from the Barrett's lesion passed the quality control filter and were included for further analysis. For samples from endoscopic biopsies from the patients presenting with gastric reflux and no Barrett's esophagus, 691 of 719 (96.10%) normal esophageal cells passed the quality control filter. I also examined the distribution of quality filtered samples across sample types and patients (Supplementary Figure A.1.6) or batch (Supplementary Figure A.1.7). While quality filtered samples were generally well distributed across patient and batch, I observed that one patient, GEN02023, had less than 5 gastric and esophageal cells represented in the quality filtered samples for further analysis. Another patient, GEN02024 had no esophageal cells represented. Unfortunately no sample remained, so it was not possible to rectify these under-represented patient tissue groups of cells. Upon examining the distribution by batch, I observed that, by design, batches C and D were the most well distributed, while batches A, B, and E were skewed toward particular tissue types. By design, Batch F consisted of only control samples and cells from the normal esophagus.

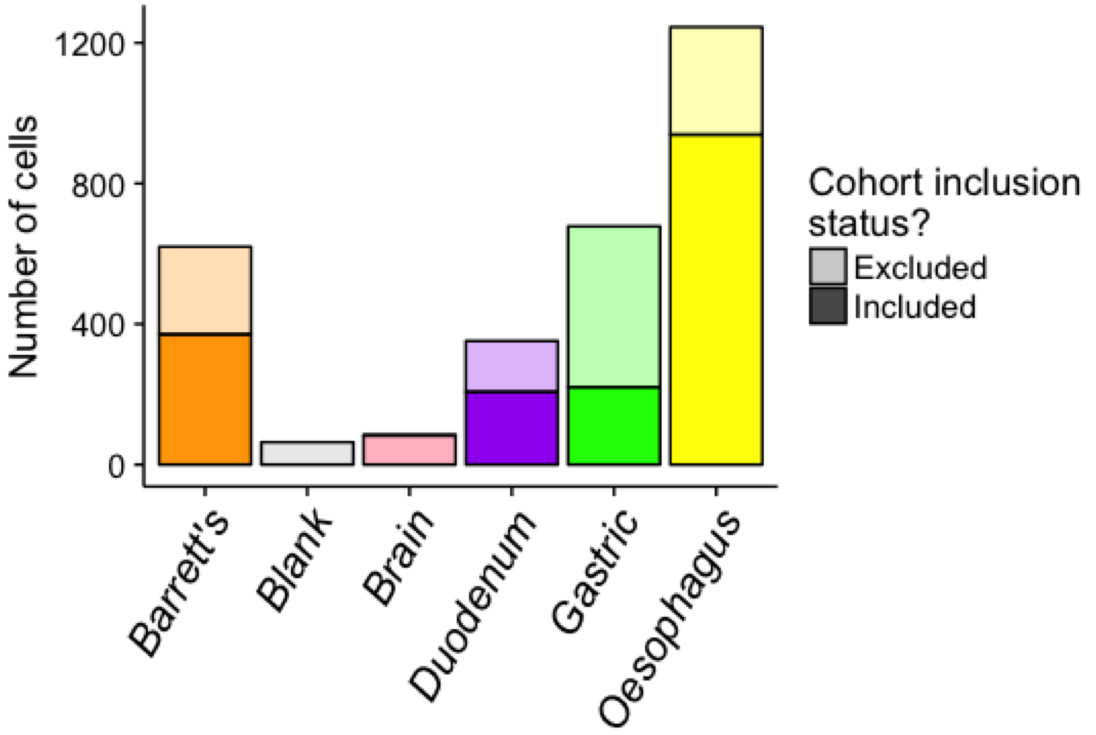


Figure 3.8.9: Barplots illustrate the final inclusion and exclusion of samples based on quality control filtering determined by total fragments mapped to the transcriptome by tissue origin (x-axis). Each bar indicates the number of cells either excluded (translucent) or included (opaque) after applying the quality control threshold. The tissue types are annotated by color.

3.9 Overview of BEARscc analysis of Barrett's experiment libraries

Once the quality control passing samples were determined, I estimated technical variation using BEARscc. Specifically, I estimated the technical variability across single cell sequencing libraries in two ways. First, because I wanted to understand the relationships amongst cell types between tissues, I estimated the noise across all tissue types including the dilute whole tissue brain control. Second, because I wanted to understand the substructure within tissue types, I estimated the noise within each tissue (esophagus, gastric, duodenum, and Barrett's) individually. One hundred simulated technical replicates were generated using the noise estimation for each individual tissue and also using the noise estimation across the entire set of samples including all tissue types.

As I observed with the murine intestinal and brain data in Chapter 2, BEARscc provides a mechanism for determining cluster number, k, using the *score* metric. However, the clustering methods utilized to cluster the murine intestinal and brain data (RaceID2 (Grün *et al.* 2016b) and BackSPIN (Zeisel *et al.* 2015), respectively) have heuristics designed to identify the best cluster number, k, for the initial clustering. In contrast, SC3 (Kiselev *et al.* 2016) provides no such heuristic and instead suggests that the SC3 consensus and the silhouette plots should be visually inspected in order to determine cluster number, k. Since I wanted to use SC3 to cluster our single cell RNA-sequencing data, I needed to determine a quantitative way to select the initial cluster number parameter, k. In order to determine the cluster number from which to cluster the observed data and the 100 simulated technical replicates with SC3, I examined the distribution of silhouette indices across simulated technical replicates for cluster numbers, $k = 1$ to $k = 15$. After using the silhouette indices to determine an initial clustering number k, I computed the noise consensus for the SC3 clustering with the aforementioned value of k for each tissue and across all cells taken together. The final cluster sets for each tissue and across all cells were chosen using the hierarchical clustering on the noise consensus and the *score* metric.

In hindsight, rather than the sequential method proposed here, which relied upon the silhouette index heuristic to establish an initial clustering parameter, a more unified approach solely relying upon the BEARscc score heuristic could have been taken; i.e. a consensus matrix would be computed for each initial cluster number k from 1 to 15, where each of the resulting 15 initial clustering k noise consensus matrices would be associated with a hierarchical clustering dendrogram split into 1 to 15 clusters as previously describe in Chapter 2. Then, the mean score metric for each clustering (15 initial k each with 15 dendrogram splits of the hierarchical clustering of the noise consensus for a total of 225 combinations) could be examined to determine the optimal pairing of the initial clustering k and noise consensus split via hierarchical clustering. I would likely take this approach if I were to reanalyze these data.

In order to further characterize the cellular subtypes within and across clusters, I performed differential expression to identify marker genes (see Section 3.3.6). Specifically, I tested for differentially expressed genes between each cluster and the rest of the data. I performed this analysis in both individual tissue types and for clusters identified across all tissue types. Next, the relationship between individual tissue types and the entire group clustering were explored. This analysis revealed a profound functional similarity between esophageal glands and ducts and the Barrett's lesion. Finally, Drs. Michael White and Richard Owen performed immunofluorescence and immunohistochemistry to validate relevant marker gene hits characterizing early markers for goblet cell types in Barrett's and the intestine and stem-like cells,

3.10 Determining clusters for biological interpretation

3.10.1 Initial SC3 cluster number parameters in individual tissues

In order to determine the working cluster number k for SC3 clustering on individual tissues, I examined the silhouette indices for clusters across the one hundred simulated replicates (Figure 3.10.1). I found that visually, the selection of the initial SC3 cluster number parameter k for gastric and esophageal tissues was obvious: $k = 3$ (mean silhouette index: 0.88 and 0.89, respectively) for gastric and esophageal tissue, respectively. However, the initial cluster number for the duodenum and Barrett's lesion required further inspection. Specifically, I was unsure whether to use k values of 4, 5, 6, or 7 for the duodenum and values 4 or 5 for the Barrett's lesion. Therefore, I computed the noise consensus and examined the distribution of cluster scores for each of these initial cluster number values for the duodenum (Figure 3.10.2) and the Barrett's lesion (Figure 3.10.3). I wanted to understand which initial SC3 cluster number parameter k had the highest possible score value amongst the lowest score values for each hierarchical clustering on the noise consensus, because I reasoned this approach would yield a set of clusters

consistently robust to noise. Therefore, I examined minimum cluster score in each hierarchical clustering on the noise consensus and compared the maximum of these values for each SC3 cluster number k. In the Barrett's lesion I found that the hierarchical split with highest scores in amongst the minimum scores within each hierarchical split were 0.61 and 0.46 for initial k values of 4 and 5, respectively. Thus, I chose $k = 4$ (mean silhouette index: 0.76) in the Barrett's lesion. In the duodenum I found the highest scores amongst the clusters with the lowest score in each hierarchical split were 0.75, 0.58, 0.73, 0.50 for k values 4, 5, 6, and 7, respectively. For the duodenum, I chose $k = 6$ (mean silhouette index: 0.67), because the scores for k values of 4 and 6 were comparable, and the duodenum is biologically expected to have diverse, specialized sub-cellular structure.

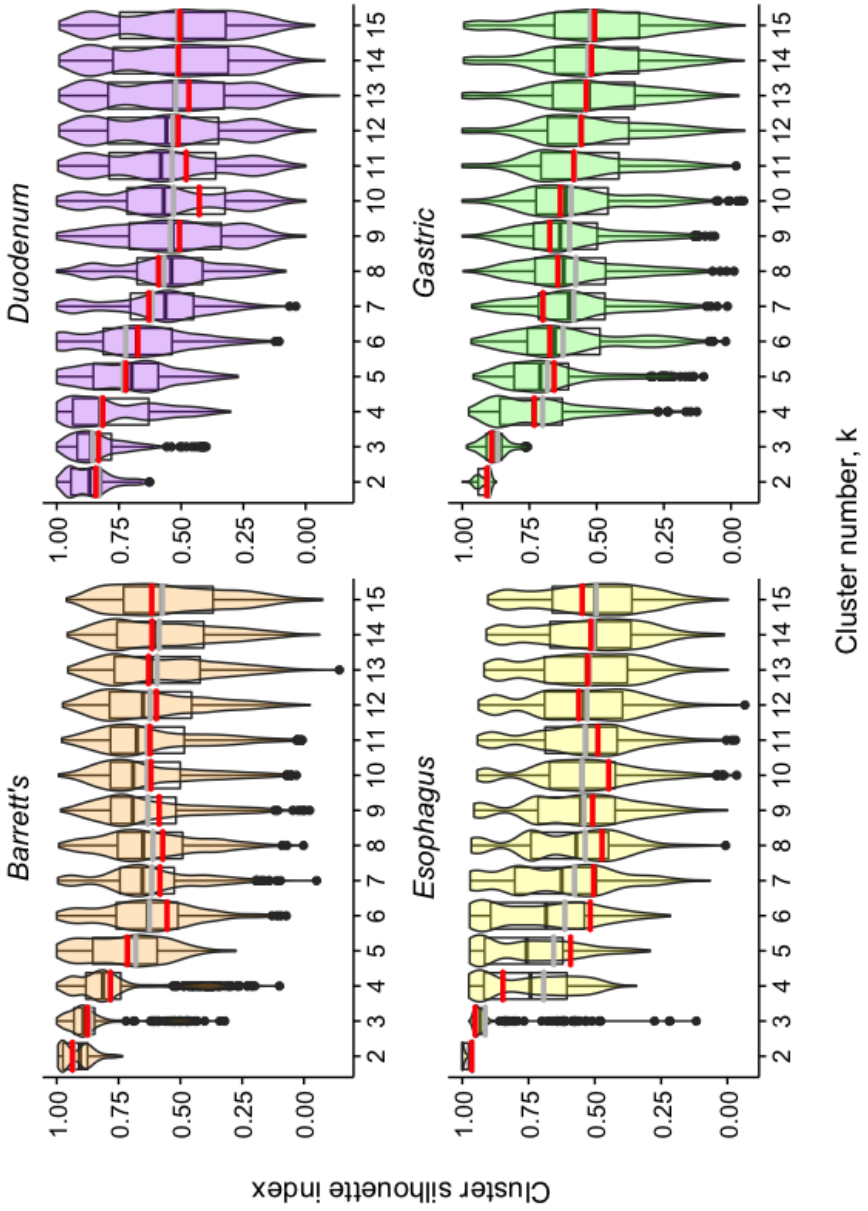


Figure 3.10.1: Boxplots overlaid by violin plots display the distribution of cluster silhouette indices across 100 noise counts perturbations by tissue type. Violin plots are colored by tissue type and indicate the distribution of samples; the width of violins is normalized across tissue or control types. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red and grey crossbars indicate the mean silhouette index for the observed and simulated data, respectively.

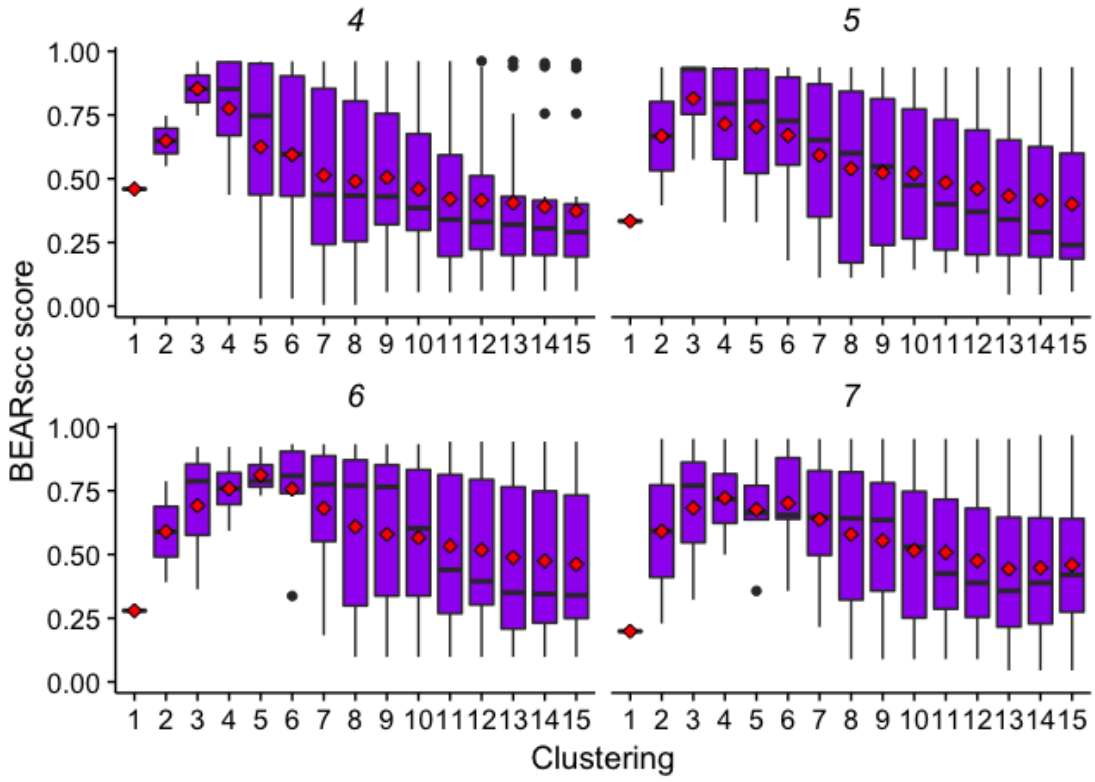


Figure 3.10.2: Boxplots display the distribution of BEARscc cluster score for each hierarchical clustering on the noise consensus of duodenal tissue at some cluster number (x-axis). Each panel is an initial SC3 cluster number parameter, where the title denotes the value of k. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

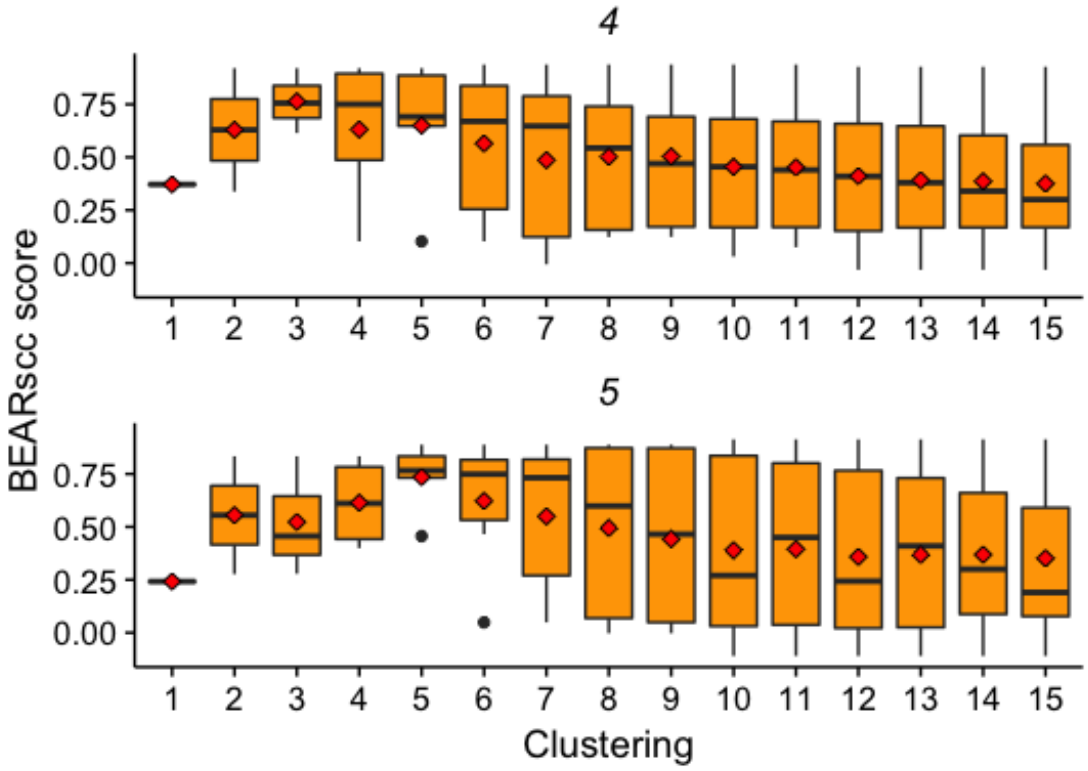


Figure 3.10.3: Boxplots display the distribution of BEARscc cluster score for each hierarchical clustering on the noise consensus of Barrett's tissue at some cluster number (x-axis). Each panel is an initial SC3 cluster number parameter, where the title denotes the value of k. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

3.10.2 Initial SC3 cluster number parameter across tissue types

Similarly, I also examined the silhouette indices for clusters across the one hundred simulated technical replicates generated from noise estimated from all tissue types (Figure 3.10.4). As with the gastric and the esophageal tissues, I found the choice for the SC3 cluster number to be visually obvious, and elected to proceed with $k = 9$ (mean silhouette index: 0.75). Indeed, I observed a local maximum amongst the mean silhouette indices at $k = 9$. Moreover, given that clustering four distinct tissues and the brain positive control into 2 or 3 clusters is not biologically sensible, then the $k = 9$ parameter also gives the highest absolute median silhouette index across simulated technical replicates amongst biologically reasonable cluster numbers.

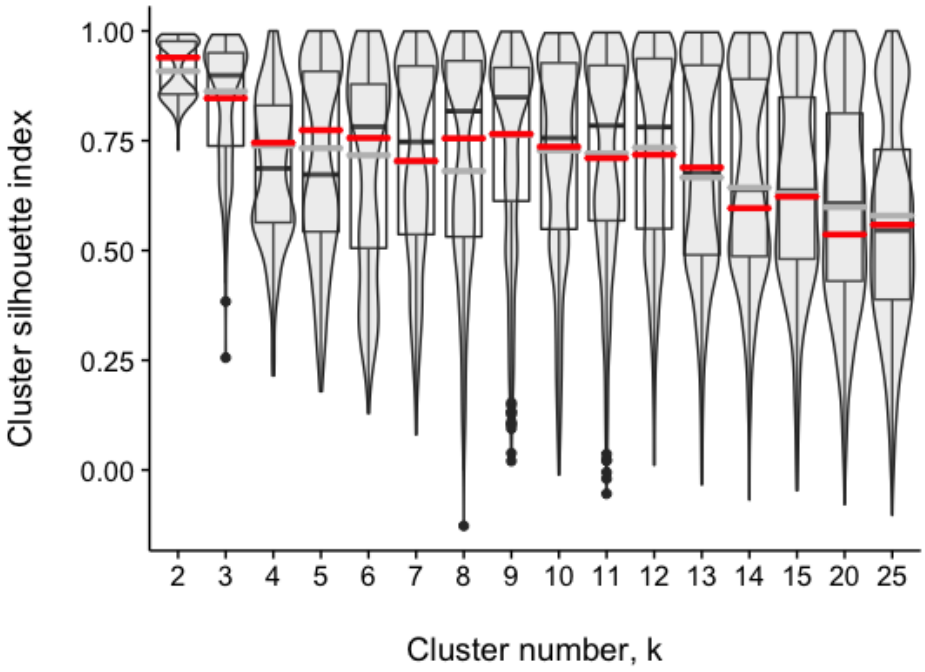


Figure 3.10.4: Boxplots overlaid by violin plots display the distribution of cluster silhouette indices across 100 noise counts perturbations. Violin plots indicate the distribution of samples; the width of violins is normalized across cluster number, k. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red and grey crossbars indicate the mean silhouette index for the observed and simulated data, respectively.

3.10.3 Final clusters are determined by BEARscc

In order to define clusters for interpretation, I examined the distributions of the BEARscc score metric on the noise consensus matrices resulting from SC3 clustering of one hundred simulated technical for each tissue with the respective chosen initial cluster number, k. The nature of the score metric has been described in Chapter 2 (Figure 3.10.5). I found that the score metric implied that there are three robust clusters in the normal esophageal, gastric, and Barrett's tissue type. I do not take this to mean that there are only three cell sub-types in each tissue, but that three clusters are well-defined and easily interpretable in the dynamic range of the single-cell experiment. In contrast, I found five robust clusters in the duodenal tissue type.

I also examined the individual components of the cluster score, stability (Figure 3.10.6) and promiscuity (Figure 3.10.7), before determining on the final cluster number to proceed with for biological interpretation of each tissue type. I found that for all three normal control tissue types, the stability clearly plateaus at the cluster number suggested as optimal by the score metric (Figure 3.10.6). In contrast, in the Barrett's clustering, the absolute maximum mean stability was observed with 5 clusters, rather than 4. Moreover, in the 5 cluster case of Barrett's esophagus, four of the five cluster were profoundly stable, where each cluster's stability metric (stabilities: 0.980, 0.978, 0.953, 0.942) was comparable to or even exceeded the maximum stability metric for the 3 cluster solution (stability: 0.953); the minimum stability metric for the 5 cluster solution (stability: 0.582) was only marginally less to the minimum stability of the 3 cluster solution (stability: 0.656). Unsurprisingly, I observed that the distribution of the promiscuity metrics were close to zero with the cluster numbers suggested as optimal by the score metric for all tissue types (Figure 3.10.7). In contrast, for any clustering solution with more clusters than the cluster number suggested as optimal by the score metric, the promiscuity distributions quickly drifted from zero towards one; this included the 5 cluster solution for the Barrett's tissue explaining the maximal score metric with 3 clusters.

As a result of examining the distribution of cluster metrics for various numbers of clusters, I elected to proceed with the cluster numbers suggested as optimal

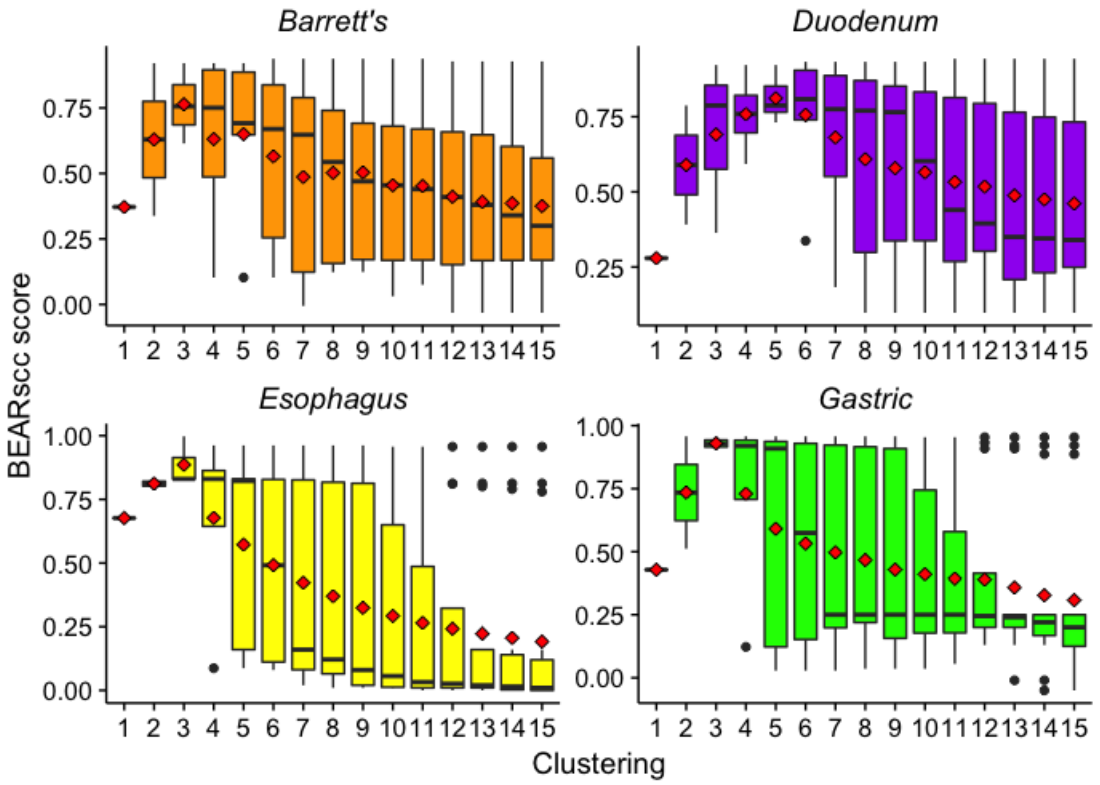


Figure 3.10.5: Boxplots display the distribution of BEARscc cluster score for each hierarchical clustering on the noise consensus of Barrett's tissue at some cluster number (x-axis). Each panel represent one of the four endoscopic biopsy sites, which are also indicated by color. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

by the cluster score for the three normal tissues. However, since lower stability scores, such as those observed in the 3 cluster solution of the Barrett's tissue, are an indication of cells within a cluster failing to associate with one another across simulated technical replicates, I reasoned that there may exist important substructure within the 3 cluster solution that is robust to noise, and that the 5 cluster solution would allow us to explore this substructure. Furthermore, our downstream analysis involved removing promiscuous cells from the dataset prior to running differential expression analysis, which would prevent the increase in promiscuity from obscuring core functional features of each cluster. Thus, I chose to

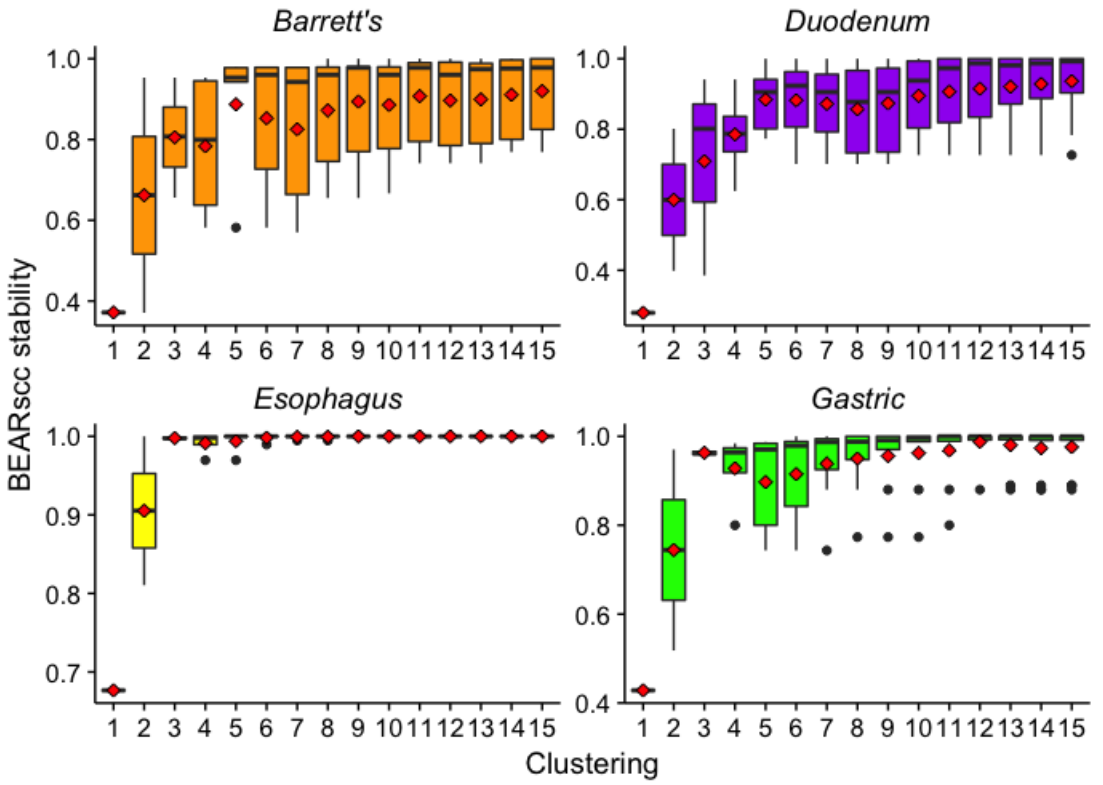


Figure 3.10.6: Boxplots display the distribution of BEARscc cluster stability metric for each hierarchical clustering on the noise consensus of Barrett's tissue at some cluster number (x-axis). Each panel represent one of the four endoscopic biopsy sites, which are also indicated by color. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

proceed with the 5 BEARscc clusters for the Barrett's tissue. In summary, I elected to proceed 3, 3, 5, and 5 BEARscc clusters for further biological interpretation of single cell samples from the normal esophagus, stomach, duodenum, and Barrett's lesion, respectively. The values of all three metrics for all four tissue types with the chosen cluster number may be found in Supplementary Table A.3.

I also examined the distributions of the BEARscc score metric on SC3 clustering of all tissue types taken together (Figure A.1.8). I observed a local maximum for the score metric with 7 BEARscc clusters. Six of the seven clusters had scores (scores: 0.972, 0.963, 0.865, 0.825, 0.822, 0.814) that were comparable to or above

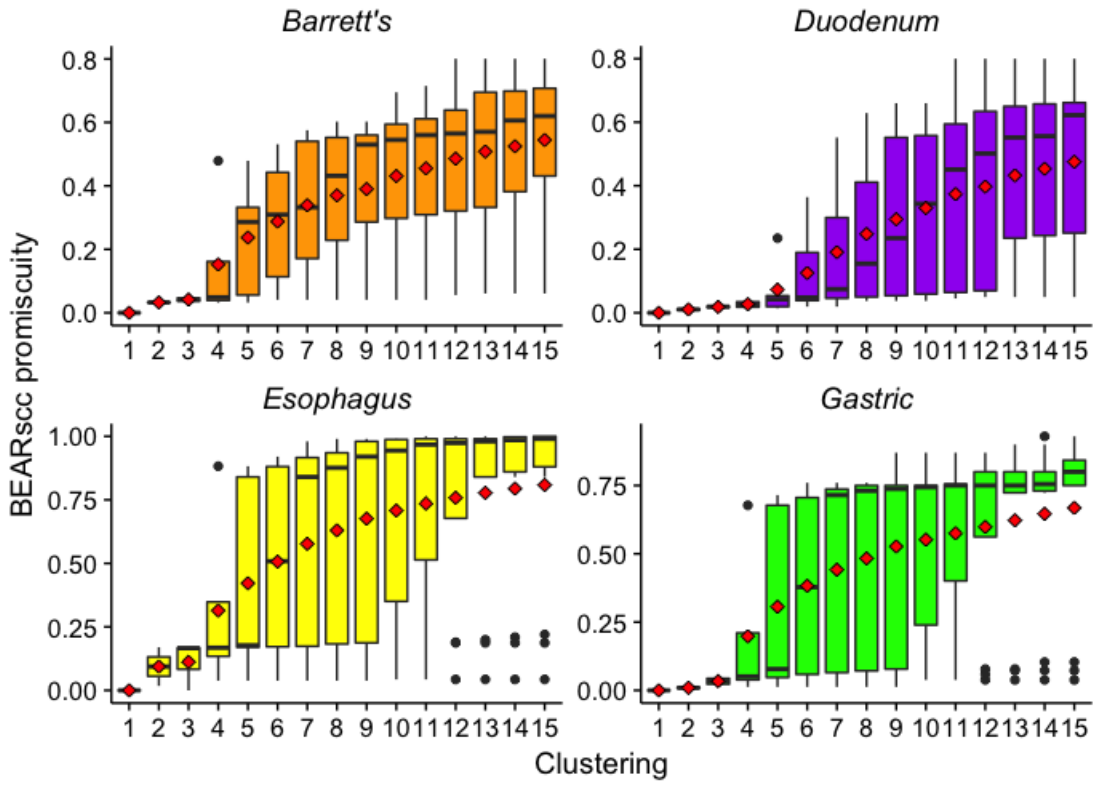


Figure 3.10.7: Boxplots display the distribution of BEARscc cluster promiscuity metric for each hierarchical clustering on the noise consensus of Barrett's tissue at some cluster number (x-axis). Each panel represent one of the four endoscopic biopsy sites, which are also indicated by color. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

the median of all other cluster number solutions (maximum median score: 0.823). Moreover, the stability distribution of the 7 cluster solution was clearly superior to the other cluster numbers, and the promiscuity of the 7 cluster solution was close to null. It should be noted that one cluster, cluster 6, had an exceptionally low score (score: 0.586) compared to the other clusters in the 7 cluster solutions. Upon examining the the other metrics, cluster 6 had a low stability, which implies that there is likely important substructure in cluster 6 to examine further. The values of all three metrics for the 7 BEARscc clusters may be found in Supplementary Table A.16. After examining the cluster metrics, I elected to proceed with seven BEARscc

cluster for biological downstream analysis. However, I resolved to examine cluster 6 more closely when interpreting the data.

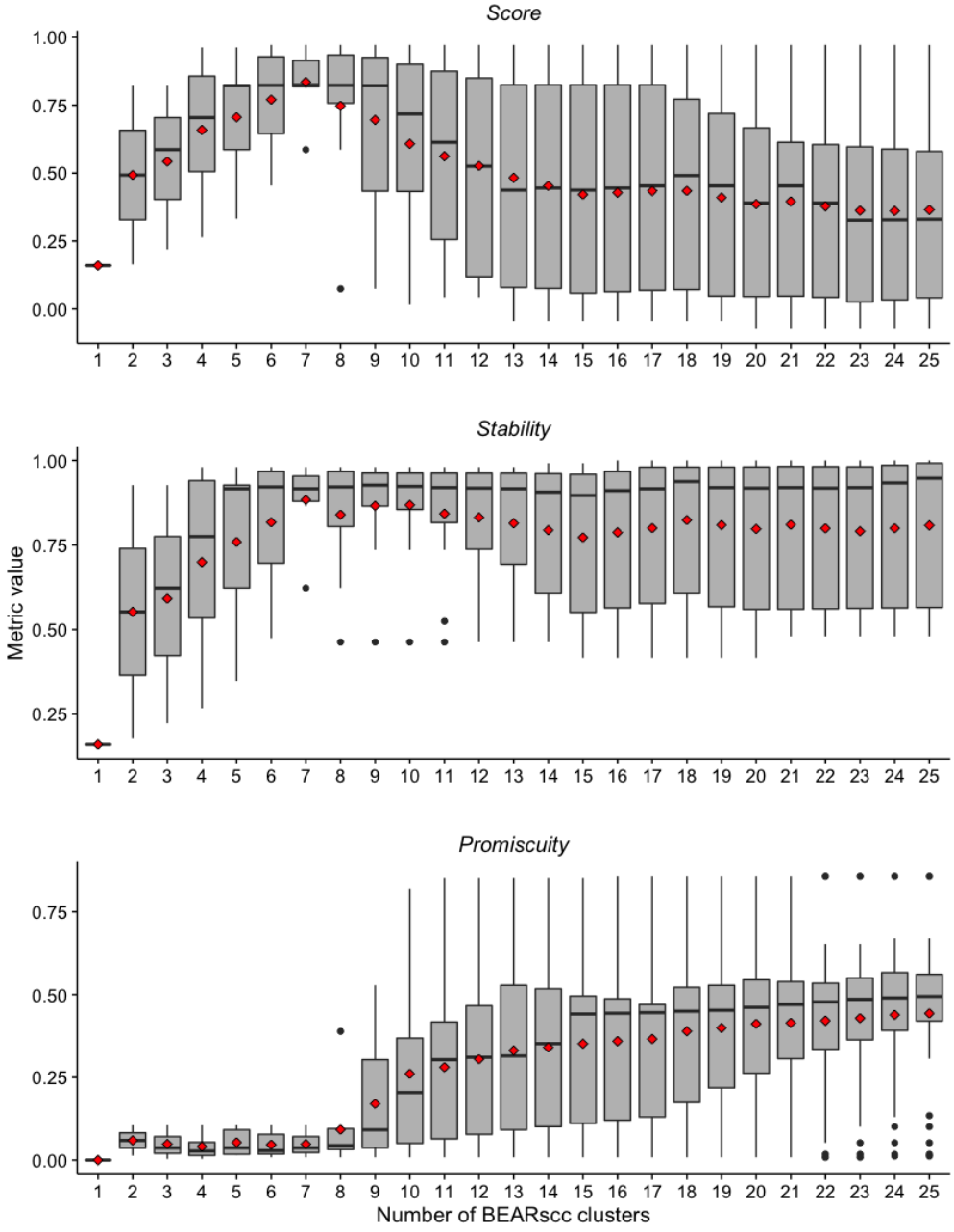


Figure 3.10.8: Boxplots display the distribution of BEARscc cluster score for each hierarchical clustering on the noise consensus of all QC-filtered samples at some cluster number (x-axis). Each panel displays one of the three BEARscc metrics: score, stability, or promiscuity. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

3.11 BEARscc identifies substructure in individual gastro-intestinal epithelia

In order to describe the transcriptional determinants of substructure in the normal upper gastro-intestinal epithelium using these single cell data, I applied BEARscc to SC3 determined clusters in cells taken from each of the individual tissue types. As previously discussed, the number of clusters initially determined by SC3 in each of the tissue types were selected using the distributions of the silhouette indices observed in various cluster numbers k (see Section 3.10.1). Final clusters were determined by performing hierarchical clustering of the BEARscc noise consensus resulting from SC3 clustering of 100 simulated technical replicates of each tissue type with the respective aformentioned cluster number, k. The number of clusters extracted from the resulting dendogram was informed by the BEARscc metrics (see Section 3.10.3).

In the supplementary, the resulting clusters for each individual tissue type are illustrated in terms of their robustness to noise, the extent to which they were confounded by patient (Supplementary Figures A.1.11[esophagus], A.1.18[stomach], A.1.33[duodenum], A.1.33[Barrett's]) and batch (Supplementary Figures A.1.12[esophagus], A.1.19[stomach], A.1.34[duodenum], A.1.34[Barrett's]), and their potential biological class. In order to biologically describe the substructure in each tissue type, I examined the noise consensus matrix for each tissue (Supplementary Figures A.1.10[esophagus], A.1.17[stomach], A.1.32[duodenum], A.1.32[Barrett's]), and determined which cells, if any, do not consistently cluster across technical replicates using BEARscc metrics (Supplementary Figures A.1.13[esophagus], A.1.20[stomach], A.1.35[duodenum], A.1.35[Barrett's]). Next, I selected the robust cells by thresholding on the cell-wise BEARscc score, and then performed differential expression analysis across clusters within each tissue type. The tables describing the chosen BEARscc metric thresholds and numbers of cells included for differential expression analysis may be found in the supplementary material (Supplementary Tables A.4[esophagus], A.5[stomach], A.6[duodenum], A.7[Barrett's]). It should be noted that further work is necessary to explore if, and to what extent, restricting differential

expression analysis to cells deemed by BEARscc to be robust to noise within their respective cluster improves the signal to noise ratio of the algorithm.

3.11.1 Summary of subcellular structure in upper GI epithelia

After obtaining differential expression results for each tissue type, I examined the resulting gene hits for each cluster and interpreted their functional class using known biology. This procedure was repeated for each individual tissue type, and a brief description of the logic of interpretation can be found in the supplemental material. Specifically, the expression of all differentially expressed genes (Supplementary Figures A.1.15[esophagus], A.1.22[stomach], A.1.37[duodenum], A.1.37[Barrett's]), and a curated list of biologically relevant differentially expressed genes (Supplementary Figures A.1.14[esophagus], A.1.21[stomach], A.1.36[duodenum], A.1.36[Barrett's]) are presented. In several tissue types, further visualization of specific gene expression was illuminating for interpretation. Therefore, boxplots of such genes will also be found in the supplemental for gastric (Supplementary Figure A.1.23[PIGR, LYZ, MUC4, LCN2, and MUC5AC genes]), duodenal (Supplementary Figures A.1.30[LGR5 gene], A.1.31[CHGA, CHGB, SCGN, REG4 genes]), and Barrett's (Supplementary Figure A.1.38[LGR5 gene]) tissues. A summary of the resulting interpretation for each cluster as well as the marker genes relevant to each interpretation are listed in Table 3.5. Note that the nomenclature of the clusters within tissue has changed for clarity, so that each numbered cluster been pre-fixed with G for gastric, B for Barrett's, E for esophagus, and D for duodenal as appropriate.

Tissue	Cluster	Marker genes	Biological interpretation	Size
Barrett's	B1	MUC2+, ITLN1+, SPINK4+	goblet	75
Barrett's	B2	CHGA+, CHGB+, SCGN+	enteroendocrine	35
Barrett's	B3	OLFM4+, HES6+	secretory progenitor	10
Barrett's	B4	OLFM4+, LGR5+	undifferentiated	201
Barrett's	B5	MKI67+, CDK1+	proliferative	50
Duodenum	D1	VIM+, JCHAIN+	tuft/mesenchymal	6
Duodenum	D2	HELLS+, VIL1+	proliferating intestinal	40
Duodenum	D3	LGR5+, REG4+	undifferentiated	94
Duodenum	D4	ALPI+, LCT+	enterocytes	47
Duodenum	D5	MUC2+, ITLN1+, SPINK4+	goblet	20
Esophagus	E1	KRT5+, TP63+, IFI44L-	basal	762
Esophagus	E2	KRT5+, IFI44L+, KRT4-	inflamed basal	89
Esophagus	E3	TFF3+, MUC6+, KRT5-	ESGDs	88
Gastric	G1	MUC5AC+, LCN2+, MUC4+	mucous (neck?)	120
Gastric	G2	MUC5AC+, MUC4-	mucous (surface?)	79
Gastric	G3	CHGA+, SCGN+, CHGB+	enteroendocrine	22

Table 3.4: The interpretation of each cluster within the individual tissue types summarized in a single table. ESGDs stand for esophageal submucosal glandular ducts.

3.11.2 Understanding the observed cellular structure

Perhaps the most notable feature of these data is the unfortunate reality that these data are somewhat insufficient in order to accomplish the task of a complete description of the three normal gastro-intestinal tract's substructure. While the total number of cells sequenced is large, I find that the number of cells in each individual tissue type, and in the gastric and duodenal tissue types in particular, are insufficient to find groups of rare cell types. Indeed, in the duodenum I observed that there was not enough power to separate two well-differentiated secretory types (enteroendocrine and goblet in D5), let alone identify groups of rare cells such as tuft cells. The rather ambiguous conclusions in the gastric tissue concerning mucous neck and surface cells would benefit greatly from further sequencing. Additionally, a subset of cells in one patient's gastric tissue exhibited signs of intestinal metaplasia further demonstrating the need for additional patients to capture inter-patient heterogenity. Ideally, at a minumum these results could provide a resource describing

substructure of the normal epithelia of the gastro-intestinal tract and Barrett's esophagus; however, at present these data mostly illustrate the utility of BEARscc and highlight known biology using a novel technique. In order to achieve the status of being a useful resource, I suspect that we would require tissue from at least 8 more patients with similar numbers of cells isolated. In fact, in order to properly explore our initial hypothesis concerning the origin of Barrett's using this technology, a more thorough sequencing of normal gastric tissue and known gastric metaplasia would do much to inform the widely held notion that Barrett's esophagus arises from the gastric cardia. Later in section 3.12, I discuss the inter-relationships between tissues with an emphasis on Barrett's and cells that appear to be esophageal submucosal glandular duct cells. In addition to sequencing more cells as done here, there are a number of experimental approaches in the form of more targeted sequencing of cells and orthogonal methods that would further support or refute the relationship described later between Barrett's and esophageal submucosal glandular duct cells. I discuss these in the conclusions and discussion (see Section 3.13) of this chapter.

Despite the lack of power, there are some new observations that are worth noting and pursuing experimentally. Of chief interest in my mind are the goblet cells of each tissue and the markers ITLN1 and SPINK4. The goblet cell is an integral part of diagnosis of Barrett's and the key feature pathologists use to ascribe the presence of intestinal metaplasia. Given this, a thorough focused study aiming to understand the differentiation of secretory progenitors into goblet cells within gastric intestinal metaplasia, Barrett's esophagus, and the normal intestine would be of clinical interest. Furthermore, an understanding of whether goblet cells in each of the aforementioned contexts are identical in nature from a whole transcriptome perspective. The data in this work (see Section 3.12) demonstrated that cells with expression profiles consistent with those we would expect from goblet cells are more alike across tissues than compared to other cell types within the tissue, which provides evidence that goblet cells in various contexts are similar in nature. However, a well-designed targeted single cell sequencing experiments would do much to cement this work. First, in order to validate the co-clustering of goblet

cells observed later in this chapter, I would propose to use FACS sort extracted tissue identified as gastric intestinal metaplasia, Barrett's esophagus, and normal intestine using a cell surface marker identifying well-differentiated goblet cells. The RNA-sequencing data from described in this dissertation could inform whether a proposed marker might be applicable to all three tissue types. This experiment would serve to verify that the single cells assigned a goblet cell identity in this study are, in fact, genuine goblet cells. It would also provide data better suited to comparing and contrasting the expression profiles of goblet cells between tissue types. Understanding the nature of the various goblet cells could provide insight into the mechanisms underlying the aberrant development of intestinal metaplasia within Barrett's. Are Barrett's goblet cells more similar to those found in the normal intestine or those found in gastric intestinal metaplasia? Or are there markers that distinguish Barrett's secretory cells from both contexts? If the variation in whole transcriptome amongst Barrett's goblet cells was greater than that seen between goblet cells across the three contexts, then this could suggest that there is a unifying differentiation pathway driving the presence of these cells in all contexts. The alternative result would be more difficult to interpret, but could provide biomarkers to distinguish anatomically displaced gastric intestinal metaplasia from genuine Barrett's. These biomarkers could be verified using RNA FISH or better still protein immuno-histochemistry. Other experiments could work to understand the role of ITLN1 and SPINK4 in goblet cell differentiation in these three contexts, and these are further discussed in the conclusions and discussion (see Section 3.13).

3.12 Interpreting the relationships between cells of different tissue types.

In order to understand further the functional relationship between Barrett's esophagus and the previously identified substructure of individual normal tissue controls, I applied BEARscc to SC3 determined clusters formed from a clustering of the entire ensemble of QC-filtered samples from the brain control, all three control tissue types, and Barrett's esophagus. The procedure leading to biological

interpretation of these BEARscc derived clusters was identical to the analysis performed on each of the individual tissue types. As before, the patient (Supplementary Figure A.1.40) and batch distributions across clusters were examined (Supplementary Figure A.1.41), and the noise consensus (Supplementary Figure A.1.39) and BEARscc metrics (Supplementary Figure A.1.42) were utilized to identify pruned clusters for differential expression analysis. A brief summary of the biological interpretation of the clusters follows, where a more in-depth justification can be found in the supplemental material. In addition to assigning biological meaning to each identified cluster, I assessed the all-sample clusters relationship to clusters identified in each tissue individually (see Sections 3.11 and 3.11.2). As a result I obsereved a strong functional relationship between the undifferentiated Barrett's cells and the esophageal submucosal glandular ducts. Finally, I observed two new marker genes, ITLN1 and SPINK4, for goblet cell types and two new marker genes, LEFTY1 and OLFM4, for stem-like cells in Barrett's esophagus.

3.12.1 The biology of the 12 BEARscc curated clusters

The tissue composition of each of the 12 curated BEARscc clusters is of interest. To that end, I examined the number of cells from each tissue within each of the 12 clusters where robust cells were either included (Figure 3.12.1) or excluded (Supplementary Figure A.1.9). The only striking difference I observed between the composition of robust cells only tissue composition and that including all cells was a reduction in columnar cell contribution to cluster 1 and the complete absence of gastric cells in cluster 7 when exclusively robust cells are included compared to the same clusters with all QC-filtered cells included. Because, I was interested in associations between cells that were robust to noise, I focus here on the tissue composition of the robust cells only.

Several of the BEARscc clusters were dominated by a single tissue type (Figure 3.12.1). I observed that cluster 1,4, and 6b were dominated by cells form esophageal tissue. In fact, cluster 4 and cluster 6b were solely composed of esophageal cells, while 95% of the robust cells in cluster 1 were from esophageal samples. Intriguingly,

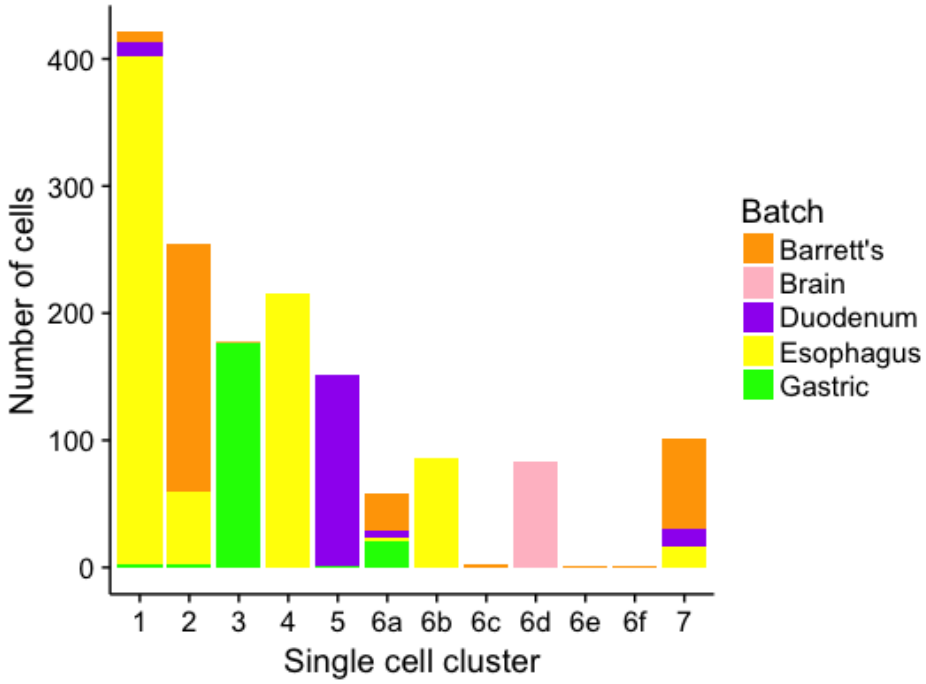


Figure 3.12.1: Some single cell clusters from all QC-filtered samples are dominated by a single tissue type and others are an admixture of tissues. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by the tissue or sample from which the single cell samples were isolated. Only samples included in differential expression analysis are displayed.

the other 5% (8 Barrett's cells, 2 gastric cells, and 11 duodenal cells) were from tissue with columnar phenotypes. Cluster 3 was almost purely composed of gastric cells with only a single Barrett's cell in the cluster. Similarly, in cluster 5, all cells were from the duodenum with the exception of 1 gastric cell. Dilute samples from the brain whole tissue positive control were the only samples in cluster 6d, and the small clusters (6c, 6e, and 6f) were contained only Barrett's cells.

The other three clusters (cluster 2, 6a, and 7) had more heterogenous tissue profiles (Figure 3.12.1). Specifically, the majority of the 254 cells in cluster 2 were from Barrett's tissue (76.38%) with a large population of cells from the proximal normal esophagus (22.44%) and 3 gastric cells (1.18%). All four tissue types were represented amongst the 58 cells in cluster 6a; the cluster contained 29 Barrett's cells (50.0%), 20 gastric cells (34.48%), 6 duodenal (10.34%), and 3 esophageal cells (5.17%). Cluster 7 consisted of 72 Barrett's cells (70.59%), 13 duodenal cells (12.75%), and 17 esophageal cells (16.67%). I suspected that these three

clusters would be the most interesting in terms of understanding the potential functional origin of Barrett's esophagus.

As aforementioned, differential expression results informed the assignment of biological meaning to the clusters. The following table summarizes the dominant tissue type (if any), marker genes,the biological meaning assigned, and the number of cells for each of the identified clusters. Additionally, the single cell libraries were visualized in two dimensions using t-stochastic neighbor embedding (t-SNE) (Krijthe 2014; Van Der Maaten 2014; van der Maaten & Hinton 2008) and annotated cells by BEARscc cluster, tissue of origin, and virtually stained cells with representative marker genes (Figure 3.12.2) to further illustrate their relationships. For reference, Supplementary Figure A.1.48 visualizes each cell's patient and tissue of origin, experimental batch, and BEARscc cluster. In summary, I identified clusters 1, 4, and 6b as basal squamous esophageal cell types with increasing degrees of stress and/or decreasing degrees of differentiation indicated by high KRT5 and stress marker KRT17(Figure 3.12.2; top row - middle and top row - right, respectively). I concluded that the columnar cells present in cluster 1 were either the result of low total gene expression or cells with non-epithelial character indicated by expression of immunoglobulin molecules such as JCHAIN (Figure 3.12.2; second row - middle). Cluster 2 was characterized by LEFTY1 expressing undifferentiated Barrett-type cells including numerous cells form the normal esophagus (Figure 3.12.2; bottom row). Cluster 3 was identified as gastric surface mucosa through expression of gastric specific genes such as GKN1 (Figure 3.12.2; third row - right), and cluster 5 consisted of undifferentiated intestinal cells (Figure 3.12.2; second row - left) and differentiated absorptive enterocytes (Figure 3.12.2; second row - right). Notably the undifferentiated Barrett's type cluster 2 also exhibited VIL1 expression. Cluster 6a cells were identified as entero-endocrinocytes by expression of marker genes such as CHGA (Figure 3.12.2; third row - middle), and cluster 6d cells purely from brain whole tissue were clearly separated with neuronal genes such as NRGN (Figure 3.12.2; top row - left). With clear secretory markers such as MUC2, I identified cluster 7 as well-differentiated goblet-type cells. Finally, the relationship

between patient, tissue, and cluster can be observed with commentary in a riverplot (Supplementary Figure A.1.49) found in the supplementary material.

Tissue	Cluster	Marker genes	Biological interpretation	Size
Esophagus	1	KRT5+, KRT13+	squamous	516
Mixed	2	LEFTY1+, OLFM4+	Barrett's-type/ESGDs	292
Gastric	3	PGC+, GKN1+	gastric mucosa	188
Esophagus	4	KRT5+, KRT13+	squamous	295
Duodenum	5	VIL1+, ALPI+	intestinal	166
Mixed	6a	CHGA+, SCGN+, VTN+	enteroendocrine	62
Esophagus	6b	LGALS7+, KRT8+	inflamed squamous	89
Barrett's	6c	TFF3+	Barrett's unexplained	3
Brain	6d	NRGN+, MOBP+	brain control	83
Barrett's	6e	TFF3+	Barrett's unexplained	1
Barrett's	6f	TFF3+	Barrett's unexplained	1
Mixed	7	MUC2+, SPINK4+, ITLN+	goblet	89

Table 3.5: The interpretation of each cluster across all QC-filtered cells in a single table. ESGDs stand for esophageal submucosal glandular ducts.

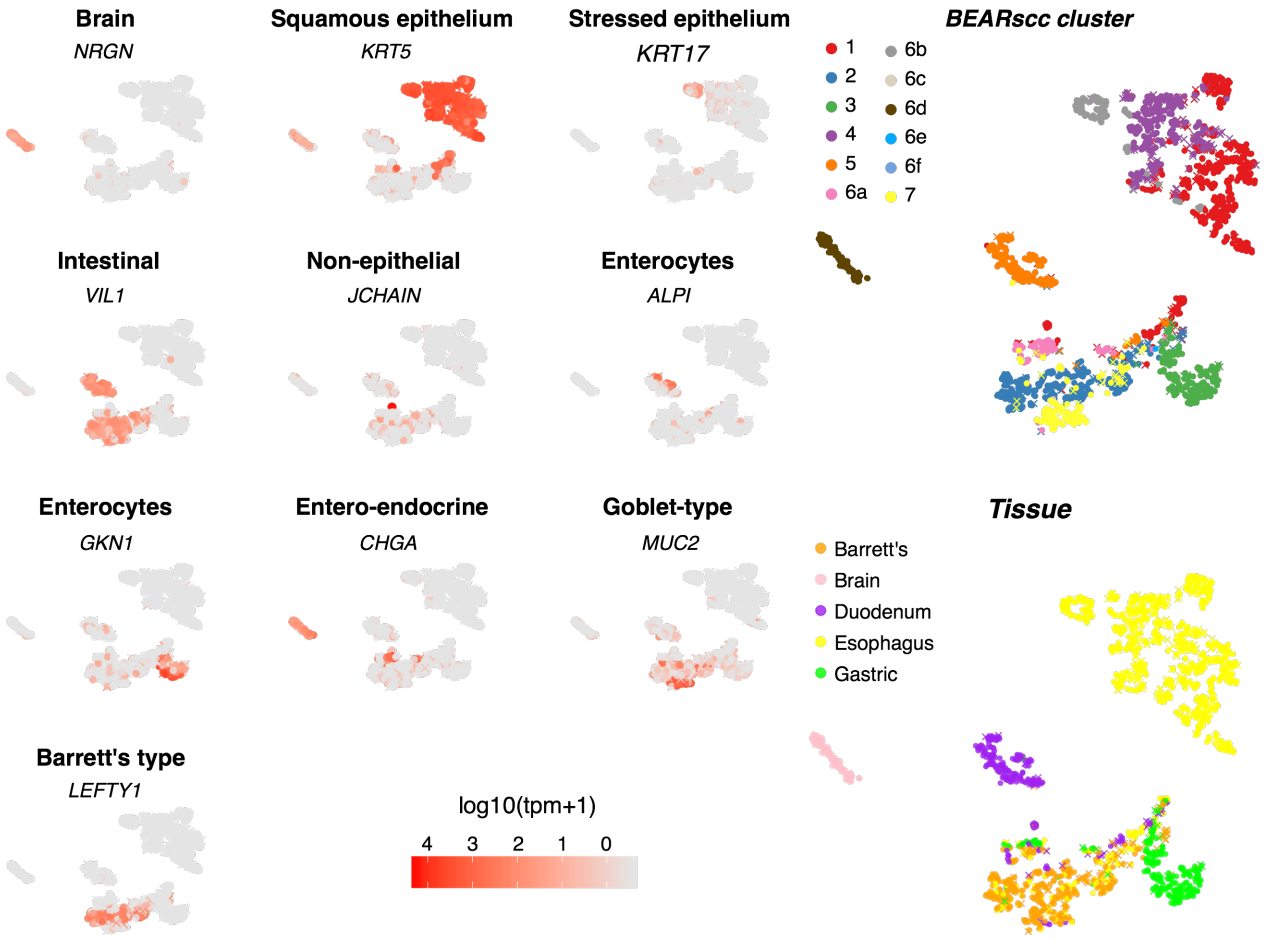


Figure 3.12.2: Two-dimensional visualizations of all QC-filtered single cells using t-stochastic neighbor embedding (t-SNE). Single cells are colored by the assigned BEARscc cluster (upper far right) and the tissue of origin (bottom far right). The 10 panels to the left display the expression in log10(TPM+1) of 10 genes differentially expressed among the BEARscc clusters. The genes were selected, because of their known biological significance. The interpreted cell type corresponding to each gene is listed in bold above each (t-SNE).

3.12.2 Barrett's esophagus is functionally similar to esophageal submucosal glandular duct cells

In order to further interpret the all-sample clustering, I examined the relationship between the BEARscc clusters found in individual tissue analysis and the BEARscc clusters found when analyzing the all tissue types together (Figure 3.12.3). The majority of the relationships between the individual tissue clustering results and the all-sample clustering reinforce the previously described interpretations of the data. Here I highlight a few of the observed results.

The entero-endocrine cells identified in Barrett's (B2) and gastric (G3) tissue converged in the overall clustering in a single entero-endocrinocyte cluster (6a). Furthermore, secretory duodenal cluster (D5) split into the two functional components I recognized in section 3.11, where the component of entero-endocrine character clustered with entero-endocrine cells form other tissue in cluster 6a and the MUC2 positive component clustered with the other goblet cells in cluster 7. I suspect that the failure to separate these two secretory lineages in the duodenum-specific clustering was the result of the small number of cells with entero-endocrine character; this hypothesis was supported by the observation that no entero-endocrine cells in D5 were identified as robust by BEARscc. The secretory progenitor cluster (B3) identified in the Barrett's specific clustering predominantly associated with the all-sample analysis goblet cell cluster (cluster 7). Unsurprisingly, the MUC2 positive Barrett's cluster (B1) were found in the all-sample analysis goblet tissue cluster (cluster 7). In both the Barrett's esophagus and the duodenum, the two undifferentiated clusters, with (clusters B5 and D2, respectively) and without proliferation (B4 and D3, respectively), merged into a single undifferentiated cluster in the all-sample analysis (clusters 2 and 5, respectively). Notably, the differentiated absorptive enterocytes (D4) found in the duodenal-specific clustering also grouped into the general intestinal-type cluster (cluster 5) in the all-sample analysis. As previously identified, the non-epithelial duodenal cluster (D1) merged into the largest squamous esophageal cluster (cluster 1) along with an assortment of columnar cells from various individual tissue clusters that were presumably low expression cells.

Perhaps the most striking observations occur in the esophagus. While the stressed / undifferentiated basal squamous esophageal cluster (cluster E2) was preserved as cluster 6b, the more differentiated esophageal cluster (cluster E1) split into two groups (clusters 1 and 2). However, the most interesting observation was the finding that the cells from the esophageal submucosal glandular ducts (ESGDs; E3) most significantly contributed to the goblet (cluster 7) and undifferentiated Barrett's (cluster 2) all-sample analysis clusters (Figures 3.12.3 and 3.12.4). Intriguingly to the ESGDs also contributed to the entero-endocrinocyte cluster (cluster 6a). As a result of this analysis, we hypothesized that this functional similarity could implicate esophageal submucosal glands. Further work is required to validate this hypothesis, especially in the context of current consensus that gastric tissue may be the origin of Barrett's esophagus.

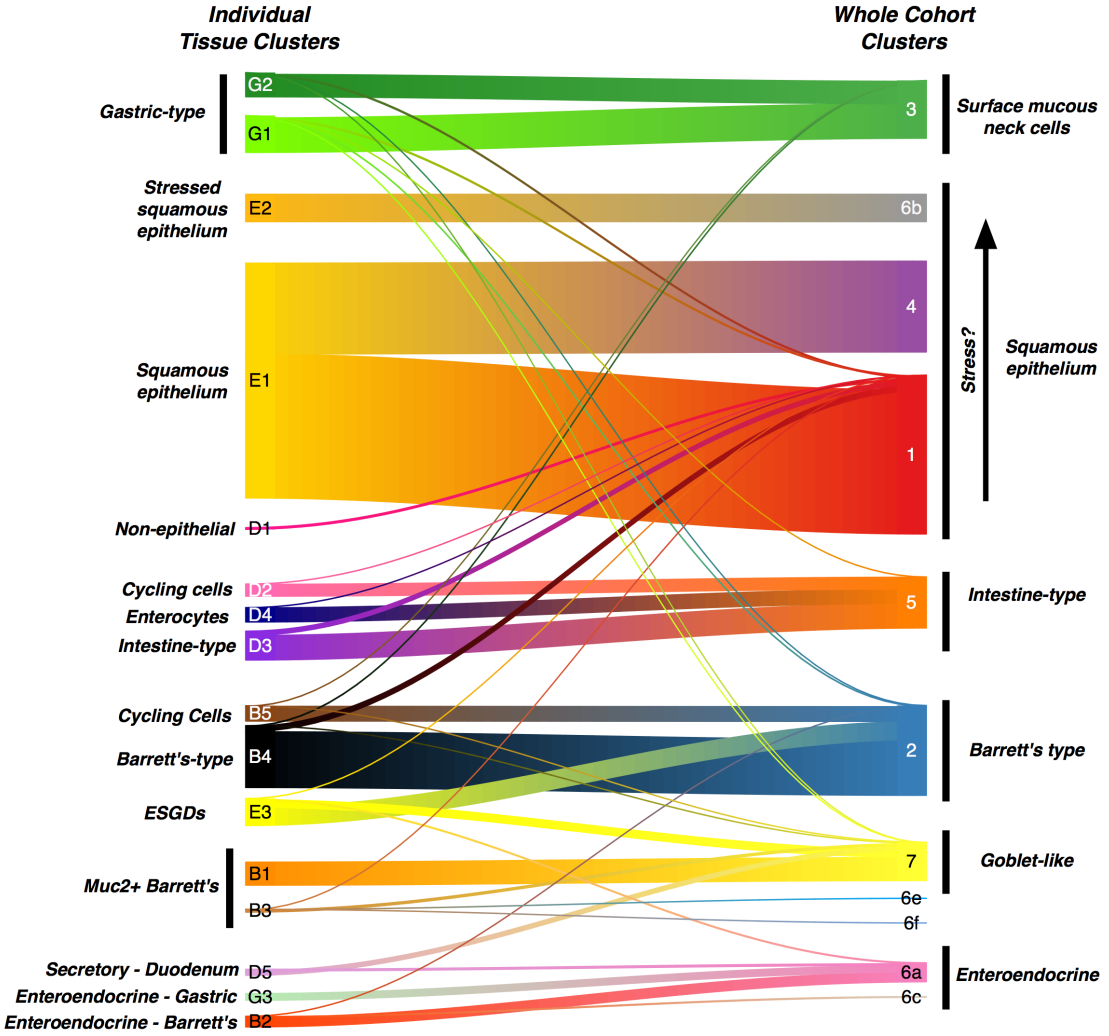


Figure 3.12.3: Sankey diagram displays relationship between the BEARscc cluster assigned in clustering of individual tissues (left) and the BEARscc clusters assigned to the sample in all-sample clustering across all QC-filtered cells (right). Nodes represent individual tissue BEARscc clusters (left) or all-sample BEARscc clusters (right), where nodes are indicated by color and node height represents the number of single cell samples in the respective calssification. Likewise edge size indicates the number of cells with shared classification between the two nodes to which it connects. Biological interpretation of BEARscc clusters is indicated in bold italics to the left of nodes for individual tissue clusters and to the right of nodes for all-sample clusters.

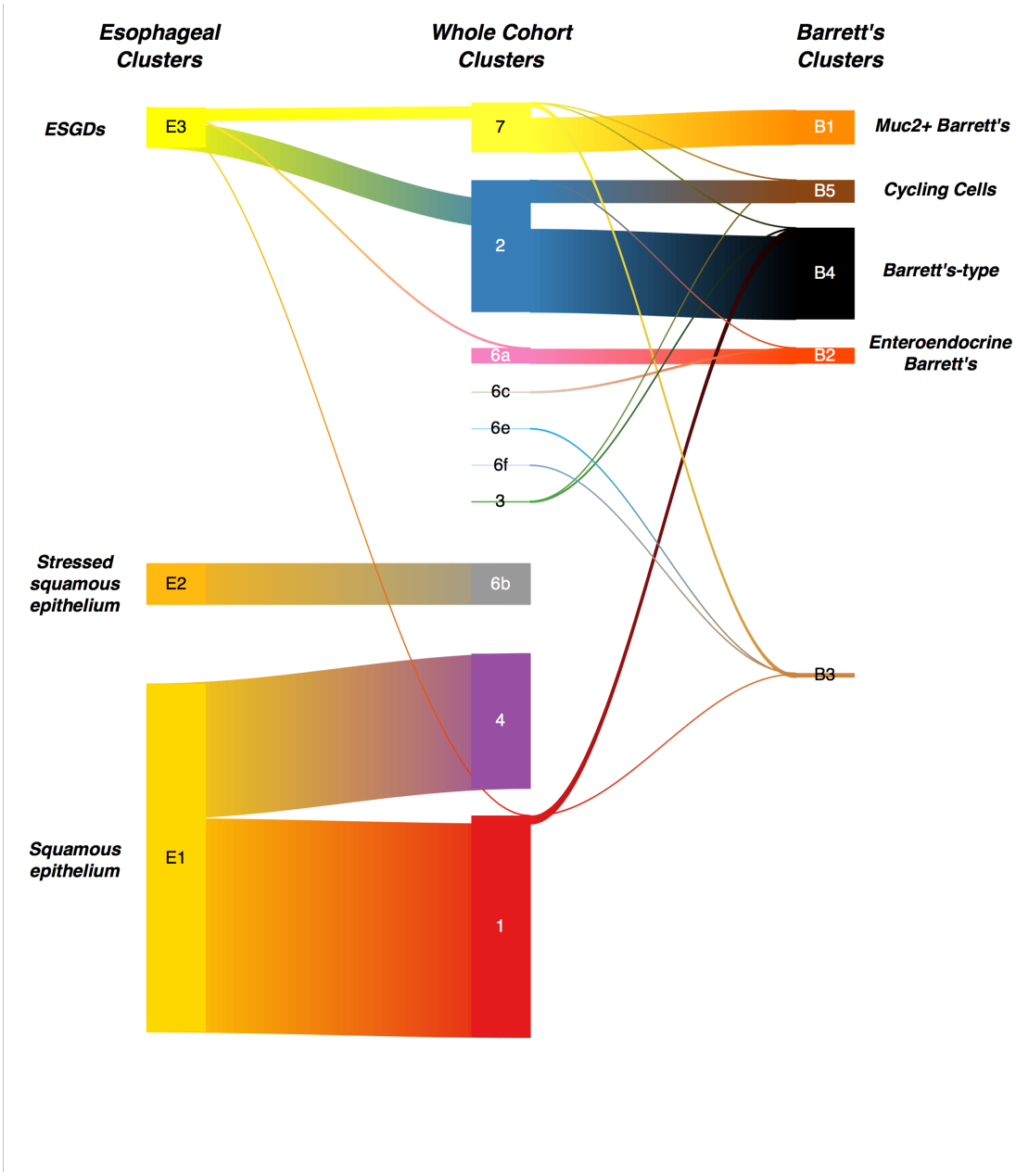


Figure 3.12.4: Sankey diagram displays relationship between the BEARscc cluster assigned in clustering of cells from esophageal tissue (left), BEARscc clusters assigned to the sample in all-sample clustering of all QC-filtered cells (middle), and BEARscc cluster assigned in clustering of samples from Barrett's tissue. Nodes represent esophageal tissue BEARscc clusters (left), all-sample BEARscc clusters (right) or Barrett's tissue BEARscc clusters (right), where nodes are indicated by color and node height represents the number of single cell samples in the respective calssification. Likewise edge size indicates the number of cells with shared classification between the two nodes to which it connects. Biological interpretation of BEARscc clusters is indicated in bold italics to the left of nodes for esophageal tissue clusters and to the right of nodes for Barrett's tissue clusters.

3.12.3 OLFM4 and LEFTY1 are expressed in Barrett's esophagus

Intrigued by the observation that Barrett's cells and esophageal submucosal glands associated strongly with one another in BEARscc cluster 2, we wanted to identify potential marker genes for this group of cells. Furthermore, cluster 2 had properties reminiscent of the undifferentiated cluster observed in the duodenum. In order to identify markers of stem-like properties, Dr. Richard Owen independently analyzed all cell libraries from Barrett's esophagus or identified as esophageal submucosal glandular duct cells using StemID (Grün *et al.* 2016a) (Supplementary Figure A.1.50c-d). As a control, the cells from the duodenum were similarly analyzed by Dr. Owen (Supplementary Figure A.1.50a-b). Comfortingly, the well-characterized intestinal stem marker, LGR5 (Barker *et al.* 2007; Sato *et al.* 2009), was among the top marker genes in the RaceID2 cluster identified by StemID as having the most "stem-ness" in our duodenal single cell libraries (Supplementary Figure A.1.50e). OLFM4, a known stem cell associated gene (van der Flier *et al.* 2009) was identified as a top marker gene in both the duodenal and Barrett's RaceID2 clusters with high "stem-ness" scores (Supplementary Figure A.1.50e).

Importantly, OLMF4 was also independently identified as significantly overexpressed gene in the undifferentiated cluster 4 in the analysis of Barrett's tissue independent of other clusters (Figure A.1.36 and Supplementary Table A.12). Likewise, LEFTY1 was identified as a significantly overexpressed gene in the analysis of all tissue together in cluster 2. Because I wanted to compare the expression of these potential markers for stemness in Barrett's, I examined the expression of LEFTY1, LGR5, and OLFM4 in two dimensions using a t-SNE projection (Krijthe 2014; Van Der Maaten 2014; van der Maaten & Hinton 2008) (Figure 3.12.5). While OLFM4 was expressed in duodenal, Barrett's, and esophageal submucosal glandular cells, LEFTY1 was restricted to Barrett's and esophageal submucosal glandular cells. Notably LGR5 was expressed in subset of OLFM4 positive cells in all three tissue types. Importantly, OLFM4 expression was higher and more

frequent in LEFTY1-positive cells when compared to the cells in the differentiated goblet and entero-endocrine clusters.

In order to validate these observations at the protein level, Dr. Michael White performed immunohistochemistry on 20 specimen's of resected Barrett's esophagus. Dr. White observed that the previously identified marker genes (LEFTY1, MUC2, and CHGA) for the Barrett's rich clusters (Clusters 2, 7, and 6a) have distinct expression patterns at the protein level (Exemplar staining in Supplementary Figure A.1.51a). Consistent with previous observations that OLFM4 associates with LGR5 positive stem cells in both normal and metaplastic intestinal contexts (van der Flier *et al.* 2009; Jang *et al.* 2015), Dr. White found that in the colon OLFM4 was localized to the base of crypts, which is the known location of stem cells (Exemplar staining in Supplementary Figure A.1.51b). In contrast, Dr. White observed that, in resections of Barrett's esophagus from 20 different patients, OLFM4 expression was not isolated to the base of crypt, a finding consistent with previous results concerning LGR5 expression in Barrett's esophagus (Lavery *et al.* 2014)(Exemplar staining in Supplementary Figure A.1.51c - top). In addition, the functional similarity identified between esophageal submucosal glandular ducts and undifferentiated Barrett's cells was supported by the presence of OLFM4 positive cells within gland structures of the esophageal submucosal glandular ducts (Exemplar staining in Supplementary Figure A.1.51c - bottom).

Taken together, these results suggest that LEFTY1 and OLFM4 are markers for undifferentiated, even stem-like, Barrett's cells, and that esophageal submucosal glandular duct cells share similar expression properties at the RNA and protein level.

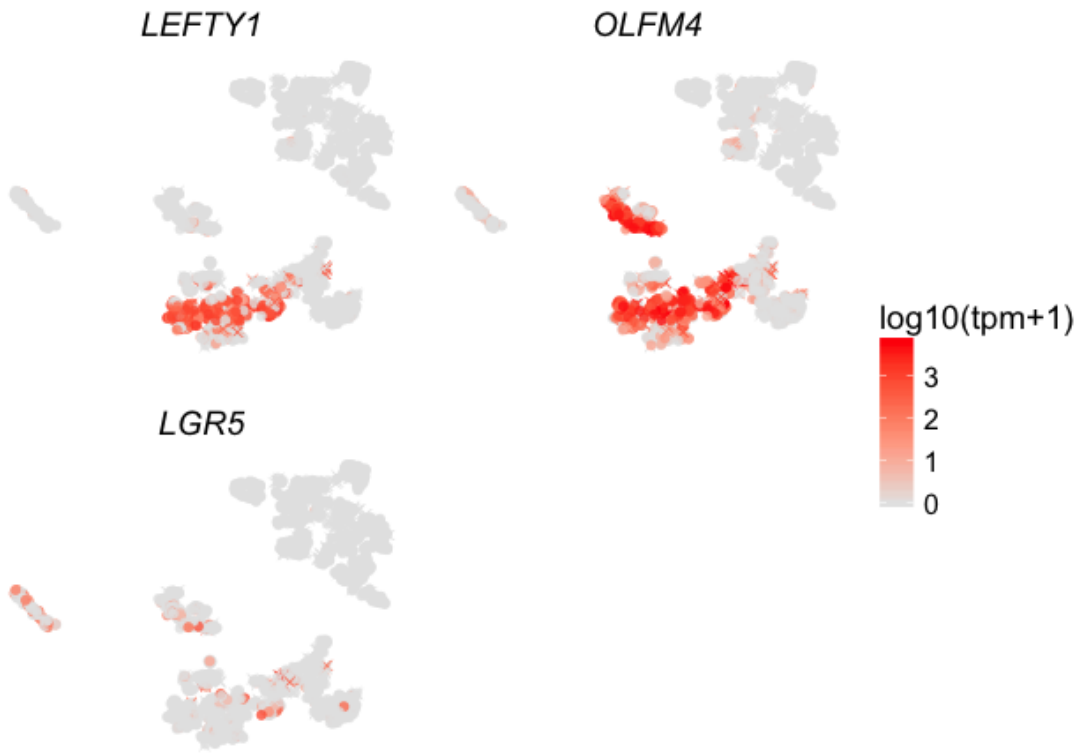


Figure 3.12.5: Two-dimensional visualizations of highlighting stem gene expression in all QC-filtered single cells using t-stochastic neighbor embedding (t-SNE). Single cells are colored by expression in log10(TPM+1) of LEFTY1 (top left), OLFM4 (top right), LGR5 (bottom).

3.12.4 ITLN1 and SPINK4 genes mark early goblet cells

In previous sections I identified ITLN1 and SPINK4 as potential marker genes for goblet cells in both individual tissue goblet cell clusters and in cluster 7 of the all-sample analysis. In order to better understand the association between MUC2 and these putative marker genes, I examined the expression of MUC2, ITLN1, and SPINK4 in 2-dimensional projection using t-stochastic neighbor embedding (t-SNE) (Krijthe 2014; Van Der Maaten 2014; van der Maaten & Hinton 2008) (Figure 3.12.6). The visualization illustrated that RNA expression of these three markers was tightly associated in the single cell data.

As previously discussed, we hypothesized that ITLN1 and especially SPINK4 may be early markers of goblet cells. In order to explore this hypothesis *in vivo*,

Dr. Michael White performed immunoflurescent microscopy on 500 samples of human colonic crypts, because colon also has a secretory lineage leading to a goblet cell fate and the differentiation of the colon is spatially well characterized with the least differentiated cells at the base of the crypt and the most differentiated cells at the top of the crypt. Dr. White observed in co-staining of ITLN1 and SPINK4, that cells expressing these markers were indeed found at the crypt base (Supplementary Figure A.1.52a). In contrast, the well-differentiated goblet marker MUC2 was present in cells located near the center and top of the crypts. These observations support the idea that ITLN1 and SPINK4 are markers of goblet cells in the earlier stages of differentiation.

After examining these markers in the well-understood colon, Dr. Michael White also examined their protein expression *in vivo* in esophageal submucosal glandular ducts found beneath the mucosa of two patients with Barrett's esophagus. In these section, Dr. White found that while ITLN1 or SPINK4 were co-expressed with MUC2 in 38% and 39% of cells, respectively (Supplementary Figure A.1.52b and Supplementary Table A.17). Intriguingly, amongst the cells with positive fluorescent staining for ITLN and SPINK4, 18% and 25% were MUC2-negative and did not exhibit obvious goblet cell morphology (Supplementary Figure A.1.52b and Supplementary Table A.17). In addition, Dr. Michael White consistently observed ITLN1 and SPINK4 positive cells without MUC2 in 20 resections of Barrett's esophagus (Supplementary Figure A.1.52c). Taken together these observations suggest that early goblet cells might be present in esophageal submucosal glandular ducts and Barrett's esophagus similar to the colon. Furthermore, these observations imply that ITLN1 and SPINK4 are potential markers for goblet-like cells, which may not be identifiable as morphologically goblet in nature. This result has diagnostic implications, because in the United States, Barrett's esophagus and subsequent monitoring for esophageal adenocarcinoma are strictly defined by the presence of goblet cells in the columnar esophagus.

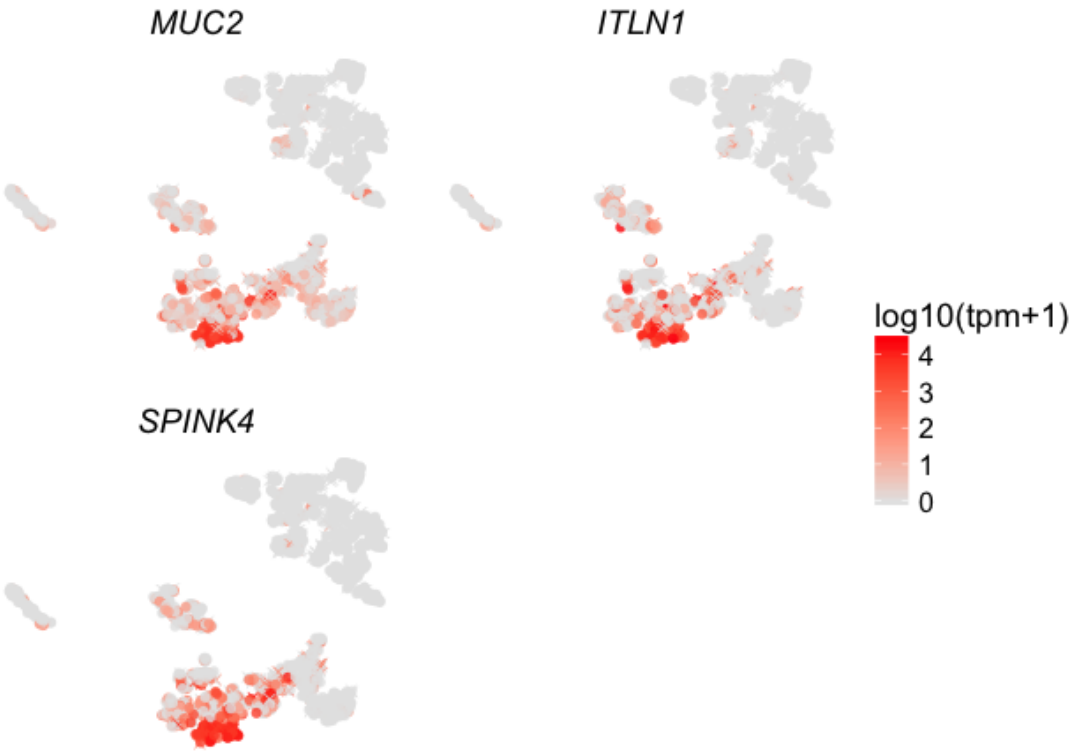


Figure 3.12.6: Two-dimensional visualizations highlighting identified goblet-type gene expression in all QC-filtered single cells using t-stochastic neighbor embedding (t-SNE). Single cells are colored by expression in log10(TPM+1) of MUC2 (top left), ITLN1 (top right), SPINK4 (bottom).

3.13 Conclusions and discussion

We successfully sequenced single cells from endoscopic biopsies of primary patient tissue in three normal tissue controls (gastric, esophageal, and duodenal) and in pre-malignant disease tissue (Barrett's esophagus). While the resulting single cell libraries generally exhibited poor mapping statistics, positive and negative control samples informed a threshold on fragments mapping to the transcriptome that isolated cells with information rich libraries with clear biologically meaningful features. In total 1821 samples isolated from endoscopic biopsies the normal esophagus, stomach, duodenum, and Barrett's lesion were successfully sequenced with sufficient quality from an attempted 3051 samples.

BEARscc analysis in conjunction with SC3 identified known biological sub-

structure of in the upper gastro-intestinal tract. Known subpopulations in the stomach, duodenum, and normal esophagus were identified. While I was able to identify sufficient substructure to provide insights into our question concerning the functional origin of Barrett's esophagus, the nuances of single cell expression in each of the three normal tissues remain to be characterized. In particular, the esophageal analysis was hindered by low sample size and confounded by genotype. Future work should aim to sequence more cells of each tissue type in order to find rare cell types, perhaps integrating a less expensive approach using 10X technologies and the macroscopic subpopulation signatures described in this work. Additionally, other clustering methods, such as RaceID2 coupled with BEARscc should be applied in order to validate the presented work and to inform details about rare cellular subtypes in each tissue. Of particular interest is the gastric tissue, where validation of the murine model of the gastric stem cell continues to be an open problem.

The cells and clusters identified by BEARscc as robust to noise provided a framework to understand the functional relationship between esophageal submucosal glandular ducts and undifferentiated Barrett's cells. By applying BEARscc to SC3, I was able to confirm that associations across tissue types were not an artifact of noise. In particular, I identified a robust cluster (all-sample analysis cluster 2) that contained predominantly undifferentiated Barrett's type cells and esophageal submucosal glandular ducts. The cluster was rich in intestinal stem cell marker LGR5 positive Barrett's and esophageal submucosal glandular duct cells. These cells also expressed the LEFTY1 and the known intestinal stem cell marker OLFM4. Critically, the expression of the shared marker genes between esophageal submucosal glandular ducts and Barrett's cells were validated *in vivo* at the protein expression level by Dr. Michael White. However, these observations do not completely exclude the possibility that esophageal submucosal glandular duct cells could in fact be the result of Barrett's tissue hidden beneath surface squamous epithelium. Although logistically challenging due to the infrequency of esophageal submucosal glandular ducts, future work should involve laser microdissection of esophageal submucosal glandular ducts from patients with no history Barrett's and patients with Barrett's

esophagus followed by single cell RNA-sequencing. This experiment would do much to affirm or refute the functional similarity between Barrett's cells and the esophageal submucosal glandular duct cells. Furthermore, the competing hypothesis that Barrett's esophagus arises from the gastric cardia is not properly addressed in this work. The number of cells obtained from gastric cardia do no provide a complete perspective for vetting this hypothesis.

It should also be noted that while these RNA-sequencing data inform similarities between tissue types, we cannot definitively address the question of origin using this experiment. However, these data could at least be used to further invigorate or detract from the notion that the esophageal submucosal glandular cells trans-differentiate into Barrett's epithelium. Using pseudo-time differentiation software, we could attempt to order the Barrett's undifferentiated cell groups (B3, B4, B5) with either the entero-endocrine (B2) or goblet (B1) cell fate. By utilizing only cells from cluster B1 or B2, we mitigate the branching problem that can confound pseudo-time algorithms. Similar analysis could be performed on the normal duodenum (with a larger sample size) as a control. We could perform identical analysis in the cells identified as esophageal submucosal glandular cells, where goblet-like cells in this tissue could serve as the terminal differentiation point. Finally, we could compare the ordering of these pseudo-time experiments and examine the genes identified as principal sources of variation relating to the ordering. Still, DNA, and not RNA, is the most likely method for identifying the origin of Barrett's esophagus. To my mind, the best evidence would arise from an experiment conducted on full resections obtained from an esophagectomy. With a full resection, distal and proximal esophageal submucosal glands, Barrett's glands, and varying degrees of metaplasia from the actual cancer could all be subjugated to targeted deep sequencing using a panel of genes previously described as being mutated in Barrett's and esophageal adenocarcinoma.

Additionally, BEARscc-SC3 clusters coupled with differential expression analysis revealed two previously unidentified markers for goblet-type cells. Further examination *in vivo* by Dr. Michael White in colonic crypts, esophageal submucosal

glandular ducts, and Barrett's esophagus confirmed that ITLN1 and SPINK4 were markers for goblet cells. Intriguingly these observations also suggested that ITLN1 and SPINK4 may be markers for early commitment toward a goblet cell lineage in all three columnar tissue types; the result has clinical implications in Barrett's diagnosis. *In silico*, goblet cell differentiation hypothesis could be explored using pseudo-time developmental ordering software as previously discussed. As proposed in section 3.11.2, a single cell sequencing approach targeting well-differentiated goblet cells in gastric intestinal metaplasia, Barrett's esophagus, and the normal duodenum could provide a clear terminal differentiation point for the three tissues. Further sequencing of isolated gastric intestinal metaplasia would also be essential. Following the collection of these data, an *in silico* experiment could be conducted for each tissue independently generating a hypothetical path of goblet cell differentiation in each tissue type. More broadly, these pseudo-time maps would ideally provide genes important in the differentiation of goblet cells in each tissue context, and these could be compared and contrasted between the three differentiation paths of the tissues. Furthermore, the expression of ITLN1 and SPINK4 could then be tracked along the three pseudo-time maps in order to describe their expression in pseudo-time. Further work *in vitro* work toward testing this hypothesis includes lineage tracing experiments with tagged ITLN1 and SPINK4 small intestine organoid clones. Following confirmation that ITLN1 and SPINK4 positive cells with no MUC2 expression preferentially differentiate into goblet cells, over-expression and knockdown experiments of ITLN1 and SPINK4 in small intestine organoids could assess whether these genes are merely biomarkers or have a functional role in delineating a goblet cell fate.

Messieurs, c'est les microbes qui auront le dernier mot.

Gentlemen, it is the microbes who will have the last word.

— Louis Pasteur

4

poshTitan: Detecting pathogens in next-generation sequencing data

Contents

4.1 Introduction **168**
4.2 Data and methods **170**
4.2.1 Gastric adenocarcinoma whole genome sequencing data . 170
4.2.2 Nasopharyngeal carcinoma RNA sequencing data 170
4.3 poshTitan workflow and implementation **170**
4.3.1 Pipeline description 170
4.3.2 Implementation 174
4.4 poshTitan accurately detects pathogens in simulated data . . . **175**
4.4.1 Performance of the Kraken fast alignment step 176
4.4.2 poshTitan is effective in simulated libraries 178
4.5 poshTitan detects *EBV* and *H. Pylori* in gastric cancer **180**
4.5.1 Landscape of pathogens in gastric cancer 181
4.5.2 poshTitan reliably detects EBV and identifies *H. Pylori* in gastric cancer 187
4.5.3 Human herpesvirus 5 (*cytomegalovirus*) was identified in two patients' tumor samples 188
4.6 poshTitan identifies EBV as ubiquitous in nasopharyngeal carcinoma **192**
4.7 Pathogens do not appear to be important in Barrett's esophagus **196**
4.7.1 No pathogen association with Barrett's esophagus is observed in whole tissue libraries 196
4.7.2 No pathogen association with Barrett's is observed in single cell libraries 200
4.8 Conclusions and discussion **204**

4.1 Introduction

Pathogens have been implicated in cancer, since the earliest days of the field of oncology (Hudnall 2014). Indeed, viruses are the primary mechanism for transformation of cells in order to establish immortalized cell lines (Ahuja *et al.* 2005; Gilbert 2001). The oncogenic risk associated with particular strains of the human papilloma virus (HPV) in cervical cancer is among the most well known examples of pathogen associated oncogenic risk (Faridi *et al.* 2011). An understanding of the importance of HPV in cervical cancer has led to vaccination efforts that have recently been shown to be successful in the reduction of cervical cancer incidence (Kim *et al.* 2016). In brain cancer, human herpesvirus 5, more commonly the cytomegalovirus, has been implicated (Söderberg-Nauclér 2015; Söderberg-Nauclér & Johnsen 2014; Wick & Platten 2014; Huse & Aldape 2014).

While the skin provides a barrier of pathogen entry with a large layer of dead cells providing a buffer, the gastro-intestinal and respiratory tracts are necessarily exposed (Donaldson *et al.* 2015). So much so in fact, that digestion is dependent on microbial activity, and the immune system is adapted to tolerate commensal microbial fauna in the gut (Donaldson *et al.* 2015). However, the presence of symbiotic commensals in the human gut and respiratory tracts certainly does not prevent infection or colonization of pathogenic viruses and microbes. Some of these pathogens are known to be oncogenic.

Indeed, the aero-digestive tract is ripe with examples of pathogens associating with oncogenic transformation. For instance, in gastric cancer the bacterium *Helicobacter pylori* has been associated with malignancy (Peek & Blaser 2002; Ye *et al.* 2004; Aituov *et al.* 2012). Specifically, the CagA expressing variants confer an increased risk for developing gastric cancer; this is true to the extent that the World Health Organization has designated CagA as an oncoprotein (Peek & Blaser 2002; Ye *et al.* 2004). Human herpesvirus 4, more commonly Epstein-Barr

virus (EBV), has been implicated in a sub-type of gastric cancer (Liu *et al.* 2015; Wang *et al.* 2014; The Cancer Genome Atlas Research Network 2014; Gulley 2015). EBV has also been frequently observed in nasopharyngeal squamous cell carcinoma (NPC) tumors and HPV is now understood to be important in a subset of oral squamous cell carcinomas (Reddout *et al.* 2007).

Barrett's esophagus is a pre-malignant lesion leading to esophageal adenocarcinoma that occurs in the lower squamous esophagus near the anatomical junction of the stomach and the esophagus. Because Barrett's esophagus occurs in exposed upper gastro-intestinal tract and given examples of cancer associations with pathogens in the stomach and nasopharynx, we wondered if there was any scope for pathogen involvement in the Barrett's lesion. In fact, the question is of general interest to the esophageal adenocarcinoma research community; there is an ongoing clinical study investigating the role of the microbiome in gastro-esophageal reflux disease and progression to esophageal adenocarcinoma (Parsons 2011). I was unable to successfully implement PathSeq (Kostic *et al.* 2011) and the scope of tools such as VirusSeq (Chen *et al.* 2013), VY-PER (Forster *et al.* 2015) and similar software is restricted to viral analysis; therefore, I elected to develop my own tool, the poshTitan, for detection of pathogens in next-generation sequencing data. In the following work, I describe the data and methods employed in this chapter (Section 4.2) including the workflow of the poshTitan (Section 4.3) and benchmark the tool using simulated libraries consisting of non-sense DNA and reads sampled from pathogens of various size (Section 4.4). I verify that the poshTitan can detect viruses and bacteria in whole genome sequencing (WGS) gastric cancer data (Section 4.5) and in RNA-sequencing nasopharyngeal squamous cell carcinoma data (Section 4.6). I then apply poshTitan to our own whole tissue and single cell sequencing libraries isolated from endoscopic biopsies taken from four tissue sites in four patients presenting with Barrett's esophagus (Section 4.7).

4.2 Data and methods

The following describes in detail the relevant data and methods employed throughout this chapter.

4.2.1 Gastric adenocarcinoma whole genome sequencing data

Professor Leung shared unmapped raw reads of published (Wang *et al.* 2014) whole genome DNA-sequencing libraries form 100 patients with gastric cancer, where matched gastric tumor and proximal normal gastric tissue were sequenced in parallel. Protocols for sample isolation, sequencing, and mapping (remaining unmapped reads were shared) are published by Wang *et al*(Wang *et al.* 2014). Unmapped reads were processed by the poshTitan as described in Section 4.3.

4.2.2 Nasopharyngeal carcinoma RNA sequencing data

Professor Ni shared raw reads of unpublished whole tissue RNA-sequencing libraries from 32 patients with nasopharyngeal squamous cell carcinoma (NPC), where matched NPC tumor and proximal normal nasopharyngeal tissue were sequenced in parallel. Sesame seed sized samples were biopsied from proximal normal and tumor tissue by Professor Ni. Total RNA was extracted by ribo-depletion and a 90bp reverse-stranded library was prepared for each sample. Sequencing was performed by BGI Genomics. Sequencing reads were processed by the poshTitan as described in Section 4.3.

4.3 poshTitan workflow and implementation

4.3.1 Pipeline description

The poshTitan is a pipeline designed to take raw read sequencing data from whole genome sequencing, single cell RNA-sequencing, and whole tissue RNA-sequencing and as input and output pathogen and commensal content in the form of binary calls with informative mapping metrics (Figure 4.3.1). The pipeline can be broken down into four steps:

1. Raw sequencing reads are filtered to remove low quality or duplicated reads and reads that align to the human genome (Figure 4.3.1a).

2. Filtered reads are checked against the NCBI viral and microbial database to flag potential hits (Figure 4.3.1b).

3. Careful realignment of candidate hits produces mapping statistics for each candidate microbe (Figure 4.3.1c).

4. Coverage statistics are used to make a binary call for microbial detection and mapping statistics are reported for further examination by the user for putative hits (Figure 4.3.1d).

In the first step, raw sequencing reads are filtered to remove low quality or duplicated reads and reads that align to the human genome (Figure 4.3.1a). Exact duplicates are removed from fastq files using a perl script written by Dr. Benjamin Schuster-Böckler:

`filter_duplicates_parallel.pl`.

The script may be found with the poshTitan source code at:

https://bitbucket.org/bsblabludwig/poshtitan.

Next, reads are hard trimmed from both the 5' and 3' end to a Illumina quality score threshold of 20 using sickle software (Joshi, N A and Fass, J N 2011), and reads with less than 32bp following trimming are discarded. The filtered and trimmed reads are then mapped to a build of hg19 including all unplaced contig files. I included unplaced contigs in the reference genome to maximize filtering of human derived fragments. For RNA-sequencing data, reads are aligned with STAR (Dobin *et al.* 2013), and the unmapped reads fastq output are saved for the next step. For whole genome sequencing data, reads are aligned using Bowtie2 (Langmead & Salzberg 2012) with the `-very-sensitive` flag parameter set in order to maximize filtering of human reads. Because Bowtie2 does not output unmapped reads as a fastq file as STAR does, for genomic data samtools software (Li *et al.*

2009) is used to filter the resulting bam file for unmapped reads and Picard tools (Wysoker *et al.* 2009) is used to dump the resulting read into fastq format.

In the second step, Kraken (Wood & Salzberg 2014) ultra-fast alignment software is used to rapidly align filtered unmapped reads to a prepared NCBI database of viruses and bacteria (Figure 4.3.1b). In order to apply the Kraken software, I downloaded all available pathogen (bacterial and viral) reference genomes from the NCBI RefSeq Nucleic Acids Research Database (National Library of Medicine US, National Center for Biotechnology Information 2002) (Access Date: August 7th 2014) and built a k-mer to lowest common ancestor database with default parameters (k-mer length set at 31 bp) as described by Wood *et al.* (Wood & Salzberg 2014). Briefly, the Kraken works by exact alignment of all possible 31 k-mers in each read fragment to the previously described database. The algorithm is extremely fast, because the database in question is loaded into memory; however, the memory demand of this step of the pipeline is extensive. The Kraken software outputs a full taxonomic classification of every read fragment in the dataset. Typically 90%-99% of the reads are reported as unclassified. I attempted to BLAST unmapped reads in a subset of samples from the cohort of gastric whole genome sequencing, but the results were uninformative. Likewise, I performed library quality analysis using FastQC (Andrews 2010) on the unclassified reads and the classified reads, but I observed no telling features in the the quality control reports that distinguished the two libraries. Thus, despite examining these reads further, I was unable to identify their source. The output of the Kraken classification is then parsed to identify the 30 pathogen *genera* with the most reads assigned to them. Admittedly, the threshold of 30 pathogens is an arbitrary cut-off, which I chose after observing that the number of fragments assigned to a genus approaches zero after approximately the first 20 hits. However, this is a tunable parameter where the optimal number is likely to be dataset dependent. These 30 pathogen *genera* are the putative positive hits designated for careful realignment by Bowtie2 (Langmead & Salzberg 2012).

In the third step, for each positive genus hit, the reads assigned by Kraken to a top genus are realigned using Bowtie2 (Langmead & Salzberg 2012) to all

relevant pathogen species' reference genomes with the `--very-sensitive` flag parameter set (Figure 4.3.1c). Prior to re-alignment, the NCBI reference genomes for all virus and pathogen species were curated into a database, where each species now has a pre-built Bowtie2 reference index. After Bowtie2 re-aligns the unmapped reads to the relevant pathogen species reference genomes, for each of the 30 genus candidate hits, the number of reads mapped to each species is computed and the identified species for each genus is taken to be the most successful alignment in terms the number of reads mapped concordantly.

In the final step, for each genus candidate hit, the mapping statistics and coverage are computed from the bam file corresponding to the species with the most successful alignment Figure 4.3.1d). Species with at least 2% coverage of their respective genome are reported as detected in the sample and the rest are considered false positive hits. More work is necessary to optimize this threshold; this is particularly true in RNA-sequencing data, where only a fraction of the genome is expressed. Finally, all positive hits above the threshold are reported along with the number of concordant alignments to the pathogen genome and the number of base pairs covered by the aligned reads. This is the final output of the poshTitan, and the following results reflect the output described in each of the respective datasets.

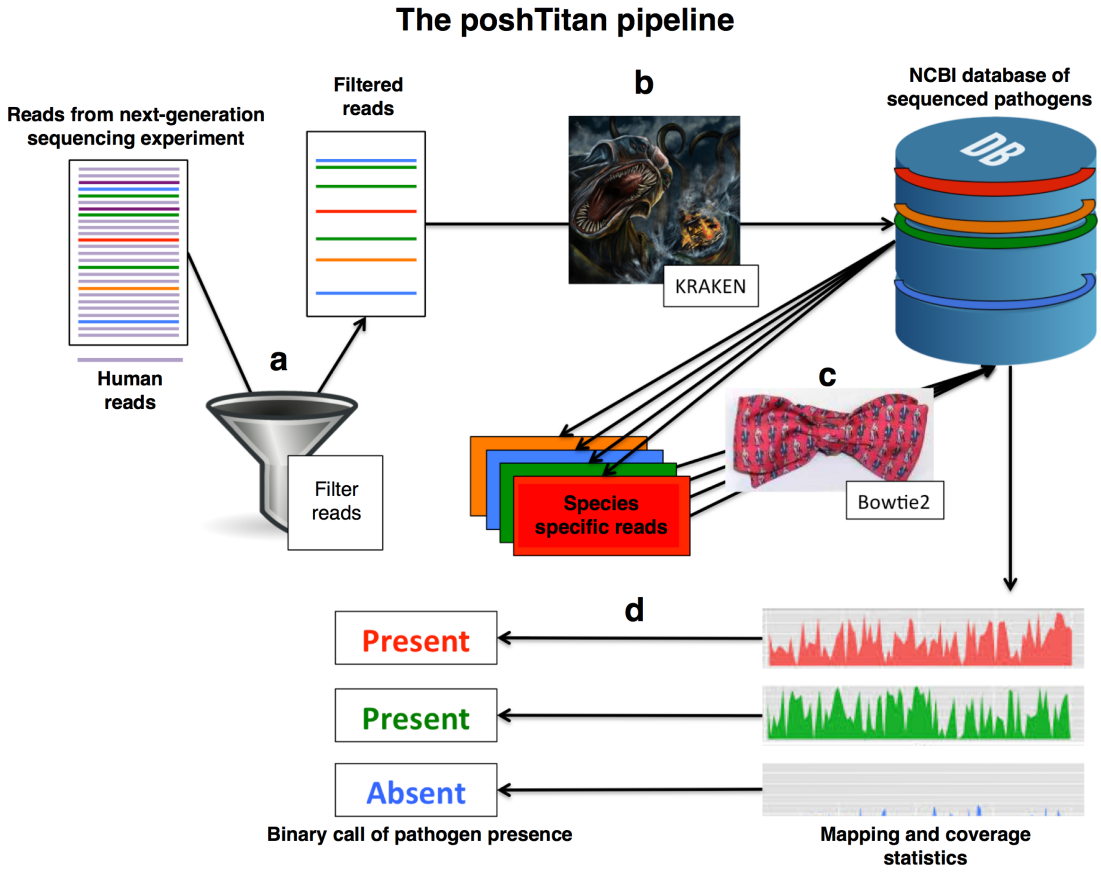


Figure 4.3.1: A schematic illustrating the flow of data in the poshTitan algorithm. Arrows denote the flow of data through the poshTitan pipeline. **a**, In the first step, raw sequencing reads are filtered to remove low quality or duplicated reads and reads that align to the human genome. **b**, In the second step, Kraken (Wood & Salzberg 2014) ultra-fast alignment software is used to rapidly align filtered unmapped reads to a prepared NCBI database of viruses and bacteria. **c**, In the third step, for each positive genus hit, the reads assigned by Kraken to a top genus are realigned using Bowtie2 (Langmead & Salzberg 2012) to all relevant pathogen species' reference genomes and summary statistics are reported for coverage and mapping from these alignments. **d**, Species with at least 2% coverage of their respective genome are reported as detected in the sample and all other species are considered false positive hits.

4.3.2 Implementation

The software, poshTitan, currently exists as a shell script pipeline with perl and R sub-scripts for file parsing, statistical analysis, and graphical reporting. The pipeline is still a work in progress, but is publicly available:

https://bitbucket.org/bsblabludwig/poshtitan.

Formulation of the code into a user-friendly and portable format suitable for wider distribution is the chief task remaining before the algorithm is complete.

4.4 poshTitan accurately detects pathogens in simulated data

In order to validate the accuracy and sensitivity of the poshTitan pipeline, I simulated 20 genomic libraries generated from random nucleotide sequences and 7 viruses and bacteria. Specifically, I simulated 20 different libraries with 4 different library compositions using Grinder (Angly *et al.* 2012), where each pathogen reference contributed a number of reads corresponding to the size of their reference genome and the total number of pathogen reads was decreased from library number 1 to library number 4 by adding in contributions of random nucleotide sequences (Figure 4.4.1). This was to assess the sensitivity of the poshTitan in terms of the number of reads from a pathogen present in a sample. Then for each of the 4 library compositions, 5 libraries were simulated with increasing rates of divergence from the reference genome ('mutation rate') to make a total of 20 simulated libraries. Specifically, in each library composition there were 5 'mutation rates', where either 1 in 1000 nucleotides (0.1%), 1 in 10 nucleotides (1%), 1 in 20 nucleotides (5%), 1 in 10 nucleotides (10%), or 1 in 4 nucleotides (25%) were randomly permuted during generation of 50bp reads from the reference genomes to form the simulated library. I varied the 'mutation rate' in order to control for varying degrees of genomic diversity in actual pathogens compared to the reference. Clearly, 25% was much higher than the diversity I would expect in real pathogen individuals (exceptionally, HIV and other retroviruses may exceed this diversity) compared to the reference genome. After libraries were simulated, they were subjected to the poshTitan pipeline as described in Section 4.3.

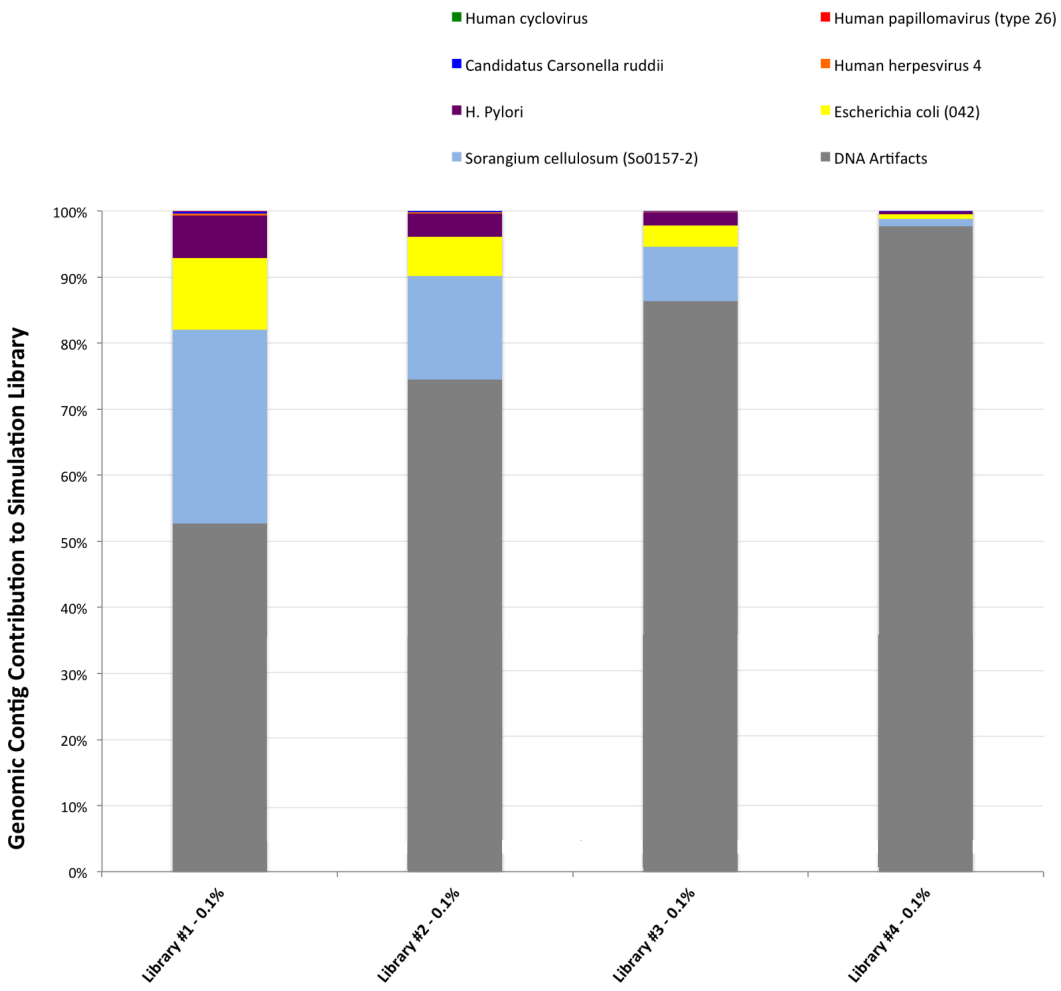


Figure 4.4.1: A stacked barplot demonstrating the construction of four libraries with decreasing pathogen and commensal contribution from exemplar reference genomes.

4.4.1 Performance of the Kraken fast alignment step

Because of the unique nature of the exact k-mer alignment algorithm in the Kraken, I wanted to specifically assess the propensity of the Kraken classification algorithm to fail to identify genuine reads sampled form reference libraries (type II error) and to assign reads incorrectly to pathogens not present in the sample (type I error). In order to assess, the propensity for the poshTitan to commit type II errors, I examined the results of processing library number 1 with the Kraken. Specifically, for each pathogen, I counted the number of reads correctly assigned to the relevant pathogen's genus and divided this number by the total number of reads

for that pathogen present in library number 1 (Figure 4.4.2). Generally, Kraken performed well correctly, detecting more than 90% of reads in 5 out of 7 pathogens for 'mutation rates' below 5%. Even at a mutation rate of 10%, at least 20% of reads were successfully classified by Kraken. Only the smallest virus, *Human cyclovirus*, failed to be identified by the Kraken, irrespective of the "mutation rate" of the library. Importantly, there were less than 20 reads present for that pathogen in the library.

In order to assess the tendency for Kraken to commit type I error, I also examined the number of reads assigned to pathogens not present in each library (Figure 4.4.3). It appeared that the majority of reads that failed to be identified by Kraken were in fact not classified at all; the false positive rate in the least accurate library (Library number 1 with a 5% 'mutation rate') was just over 0.5% or 1 in 200 reads. However, when considering the number of reads in a library may exceed several thousand, this false positive rate is worrying. This observation motivated the importance of the downstream re-alignment of putative reads to their respective candidate species.

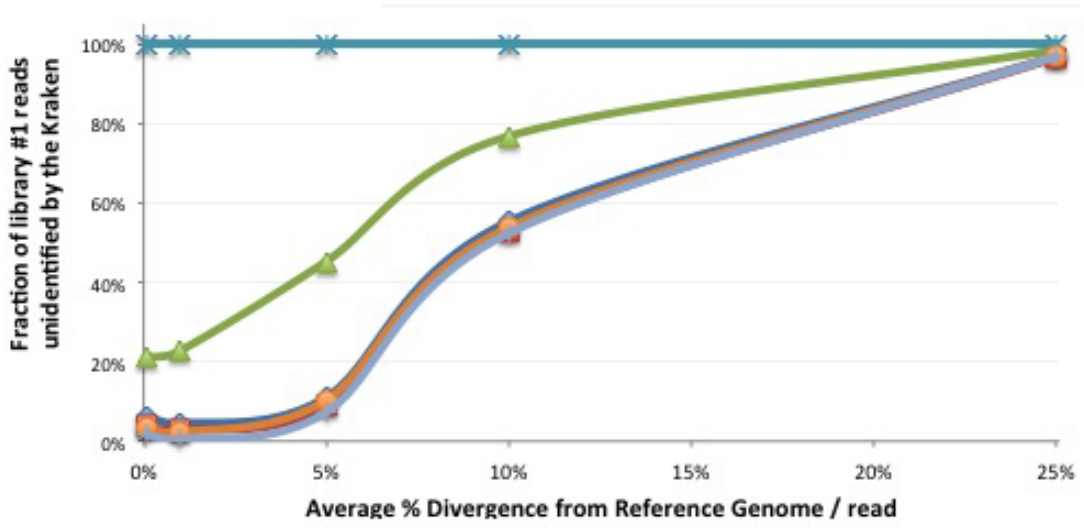


Figure 4.4.2: Scatterplot of false negative rate vs the divergence from the reference genome for library 1 with each line being one of the pathogens contributing to the simulated library: *Human cyclovirus* (aqua blue), *Esherichia coli* (green), *Candidatus Carsonella ruddi*(sky blue), *Helicobacter pylori* (orange), *Human herpesvirus 4* or *EBV* (navy blue), *Soriangium cellulosum* (red)

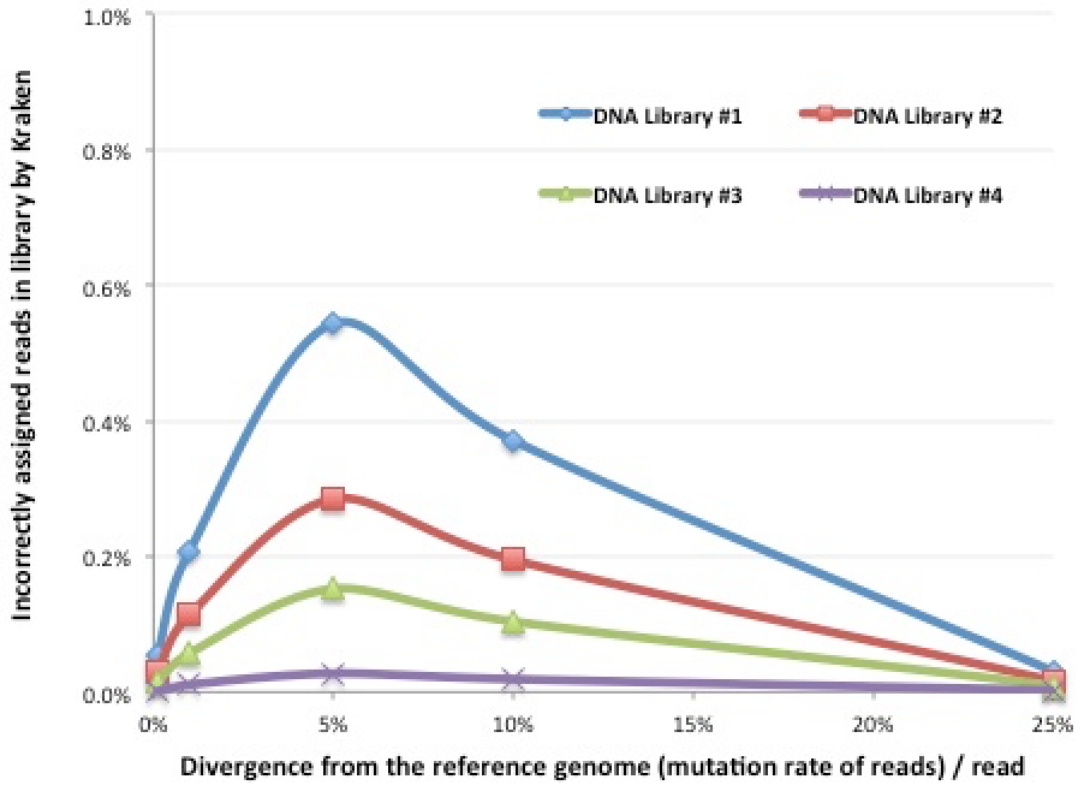


Figure 4.4.3: Scatterplot demonstrating the number of reads misassigned by the Kraken for each of the 20 simulated libraries.

4.4.2 poshTitan is effective in simulated libraries

After assessing the performance of Kraken individually on 20 simulated libraries I evaluated the performance of the entire poshTitan in terms of correctly calling pathogens from the simulated data. However, I wanted to identify a reasonable threshold for excluding false positives, so I examined all successful hits identified by the poshTitan. The poshTitan was generally accurate and sensitive: all pathogens except the *Human cyclovirus* were detected in 18 out of 20 simulated libraries (Figure 4.4.4). I observed that the principle limiting factor in sensitivity was the number of pathogen reads present in the library; sensitivity did not vary with 'mutation rate', but in the library composition consisting of approximately 100,000 reads, the two pathogens with the smallest genomes, *Candidatus Carsonella ruddii* and *Human herpesvirus 4 (EBV)* were not detected. Furthermore, the smallest viral genome, *Human cyclovirus* was not detected in any samples. The only false

positive detected was *Esherichia fergusoni*, a member of the same genus as *E. coli*. Notably the false positive occurred in the 25% 'mutation rate' library with the most pathogen reads and thus was statistically most likely to produce a reads corresponding to a closely related organism, as I observed here. Upon examining the coverage I observed that *Escherichia fergusoni* had coverage below 2% and very few mapped reads (Figure 4.4.5). Because the number of mapped reads varied greatly depending on the number of reads present in the sample, I elected to use a coverage threshold of 2% for all proceeding analyses in order to exclude false positive hits, such as the one observed here. It should be noted that there is certainly scope for further optimization of this threshold, and this will be important for future work involving the poshTitan.

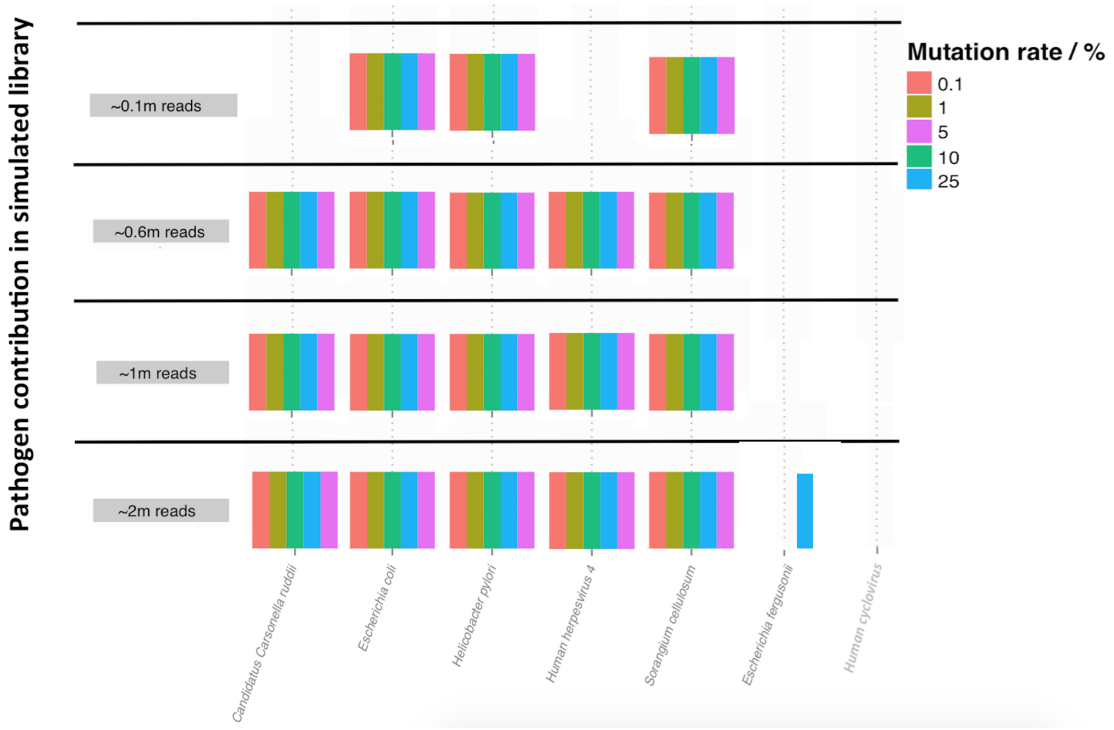


Figure 4.4.4: Barplot illustrates the pathogens and commensals detected in the simulated libraries. The presence of a bar indicates that the pathogen (x-axis) was detected in the respective library (vertical panels and color annotation). Library 'mutation rate' is indicated by color, and the total number of pathogen reads contributing to the library are separated by vertical panels.

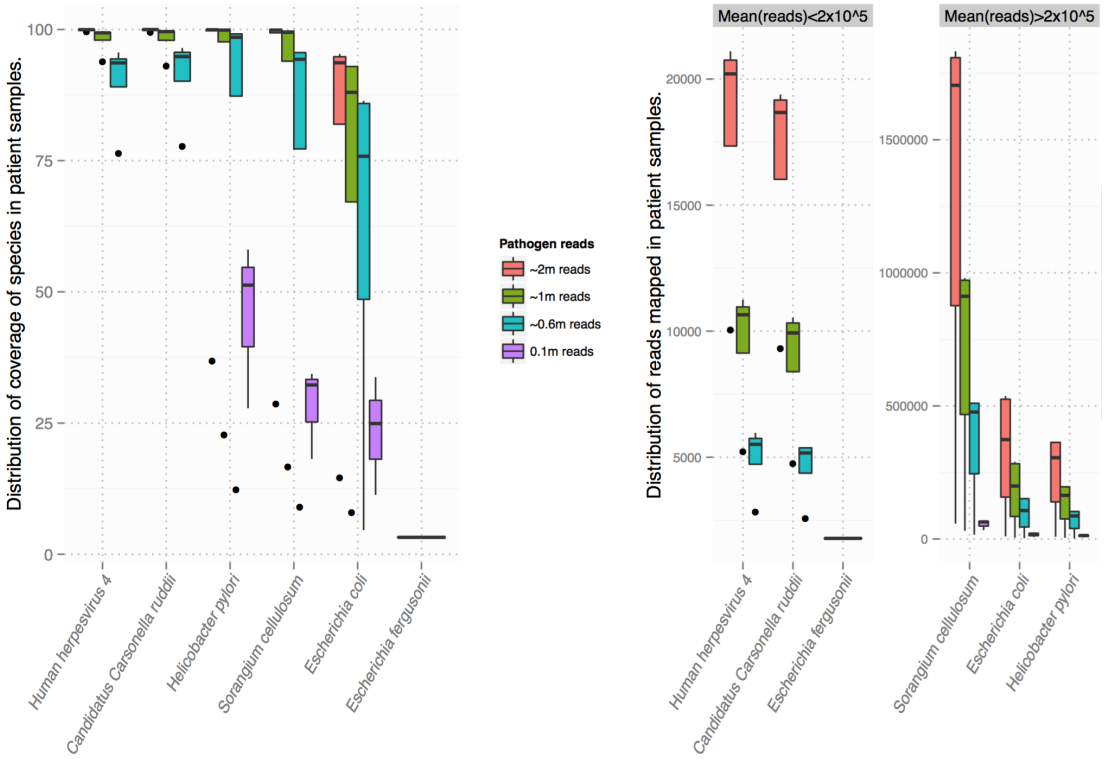


Figure 4.4.5: Boxplots illustrate the distribution of coverage (left) and mapped reads (right) for each species (x-axis) present in the 20 simulated libraries. Distributions consist of the 5 different 'mutation rates' for each of the four settings of total pathogen read content (annotated by color).

4.5 poshTitan detects *EBV* and *H. Pylori* in gastric cancer

Next I assessed the performance of poshTitan in the context of actual whole genome sequencing data. Furthermore, I tested its performance in data where pathogens were known to be important. Gastric cancer is known to be associated with both bacterial (*Helicobacter pylori*) and viral (*EBV*) pathogens. We collaborated with Professor Suet Yi Leung (see Section 4.2.1) and analyzed the unmapped reads from matched tumor and proximal normal gastric tissue from 100 patients with poshTitan. Although the reads I received were unmapped, I still performed the human reads filtering step in the first step of poshTitan.

4.5.1 Landscape of pathogens in gastric cancer

I detected a large number of pathogens and commensals with poshTitan in the gastric cancer and proximal normal samples (Figures 4.5.1 and 4.5.2). Among the most frequently detected species was the bacteriophage *Enterobacteria phage phiX174 sensu lato*. The bacteriophage *phiX174* was used as a library loading control, and thus was expected and serves as positive control for the poshTitan. Notably, where *phiX174* was detected, I observed nearly 100% coverage (Figure 4.5.3) and among the highest number of mapped reads (Figure 4.5.4). However, *phiX174* was detected in only 71 tumor and 75 normal samples out of the 100 samples analyzed of each. Given the clear results of the positive samples, I suspected that this was either the result of a change in protocol or the depth of the remaining samples was confounding the results. In any case, the observation highlighted the need to incorporate sequencing depth into poshTitan's framework in future work. Although the stomach is expected to be sterile, I observed *E. coli* in 75 normal samples and 72 tumor samples. I observed that only two (pfg058G and pfg058T) of the *E. coli* positive samples were negative for the bacteriophage. Furthermore, only one bacteriophage positive sample did not contain *E. coli*. When interpreting microbial and viral findings, it is important to consider the common phenomenon of commensalism: organisms benefit from associating with the human aero-digestive while providing no benefit nor causing harm to the human host. In line with this consideration, I also observed a number of species from *genera* associated with commensalism including *Staphlyococcus*, *Lactobacillus*, *Enterococcus*, and *Streptococcus*. *Treponema denticola* was detected in a handful of tumor and normal samples. As a common commensal in the oral cavity, it is conceivable that these bacteria contaminated those patients whose samples were retrieved by endoscopy, though that is speculation. The *Vicia cryptic virus* was observed in a number of samples, but had extremely low coverage near the 2% threshold in all positive hits (Figure 4.5.3).

Importantly, it has been observed in controlled studies that metagenomic analysis is easily contaminated by trace DNA from laboratory protocol kits used to isolate

and prepare samples for sequencing (Salter *et al.* 2014; Hemme *et al.* 2015). In their work, Salter *et al* and Hemme *et al* outline common contaminants, which typically are associated with soil and water associated *genera* (Salter *et al.* 2014). Among the *genera* identified as regular contaminants are many of the positive hits reported by poshTitan: *Propionobacterium*, *Prevotella*, *Micrococcus*, *Streptococcus*, *Fusobacterium*, *Methylobacterium*, *Ralstonia*. In fact, some of these bacteria were identified in numerous samples. Therefore, any conclusions surrounding these bacteria should be avoided. In general, while coverage provides a decent proxy for legitimate presence in the sample because contaminants are typically low-abundance, sufficiently deep sequencing of a contaminant will likely yield high genome coverage and poshTitan will call a false positive; therefore, this work is primarily observational and any interesting results should be explored further with proper controls.

I observed both *EBV* and *H. Pylori* positive samples in the gastric data. Specifically, 7 tumor samples and 3 normal samples (corresponding to 3 of the 7 postive tumor samples) were positive for *EBV*. Furthermore, following the bacteriophage, *EBV* was the second most covered pathogen genome (Figure 4.5.3) and had the second highest number of mapped reads (Figure 4.5.4). I observed that *H. Pylori* was the fourth most common pathogen with 32 patients (either tumor, normal, or both) positive for *H. Pylori* as assayed by poshTitan. In addition to *EBV*, I observed two other herpesviruses: *Human herpesvirus 7* and *Human herpesvirus 5* (*cytomegalovirus*).

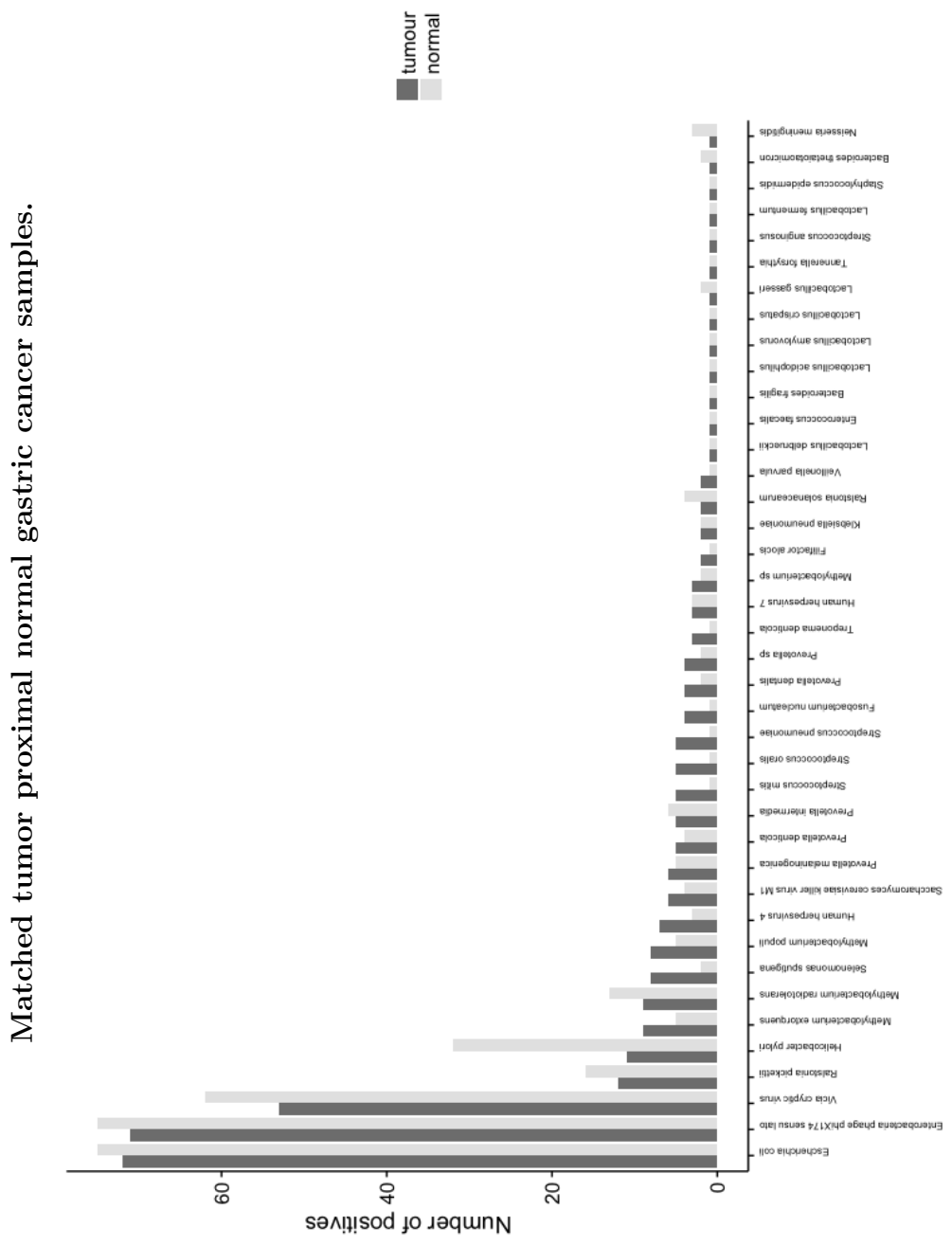


Figure 4.5.1: Barplot depicts the pathogens and commensals detected in both gastric tumors and proximal normal gastric tissue samples by poshTitan. The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) for tumor (dark grey) and normal (light grey) samples.

Matched tumor proximal normal gastric cancer samples.

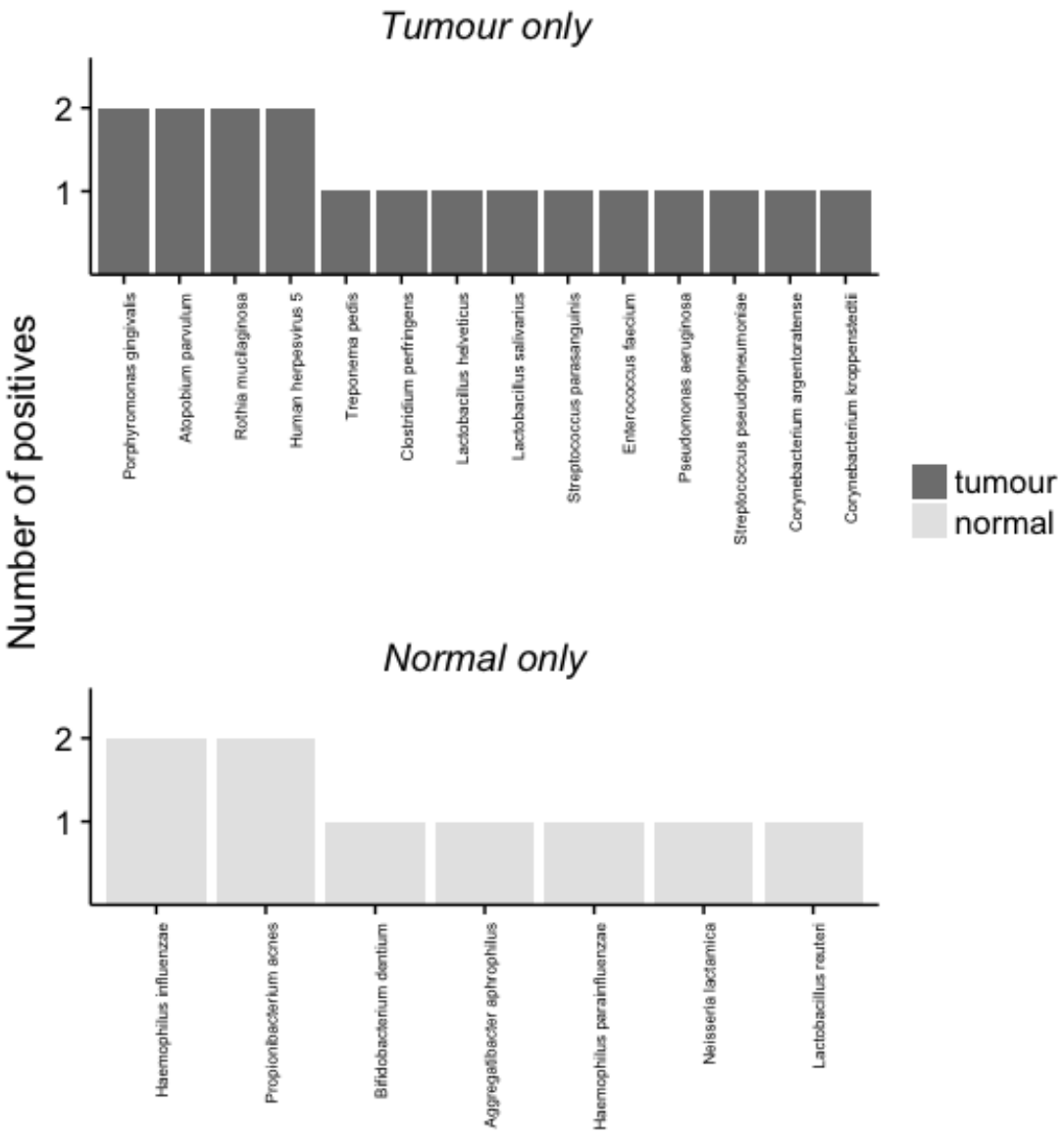


Figure 4.5.2: Barplot depicts the pathogens and commensals detected in either gastric tumor or proximal normal gastric tissue only by poshTitan. The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) in either tumor (dark grey, top) or normal (light grey, bottom) samples

Coverage of detected pathogens in gastric cancer samples.

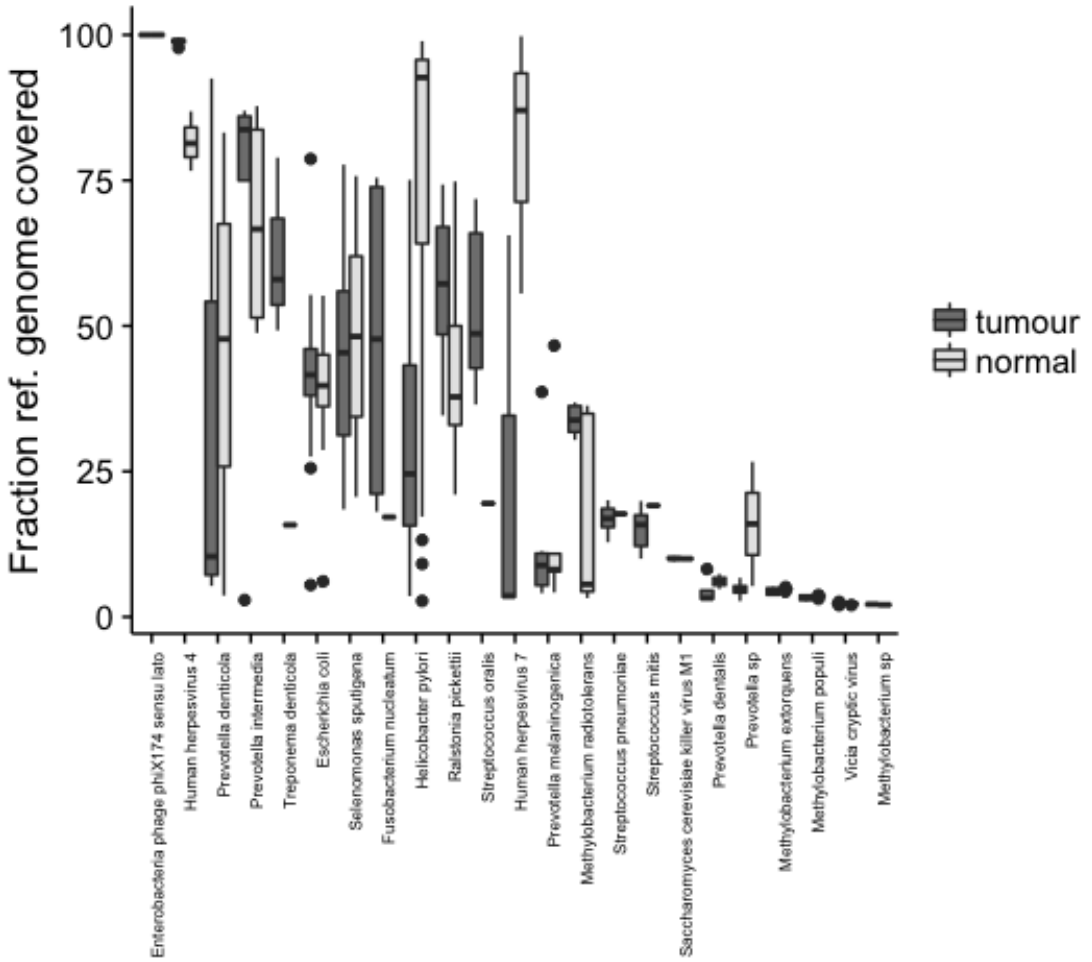


Figure 4.5.3: Boxplot illustrates the distribution of coverage for each species (x-axis) detected by the poshTitan in gastric cancer libraries. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

Number of mapped reads to pathogens in gastric cancer samples.

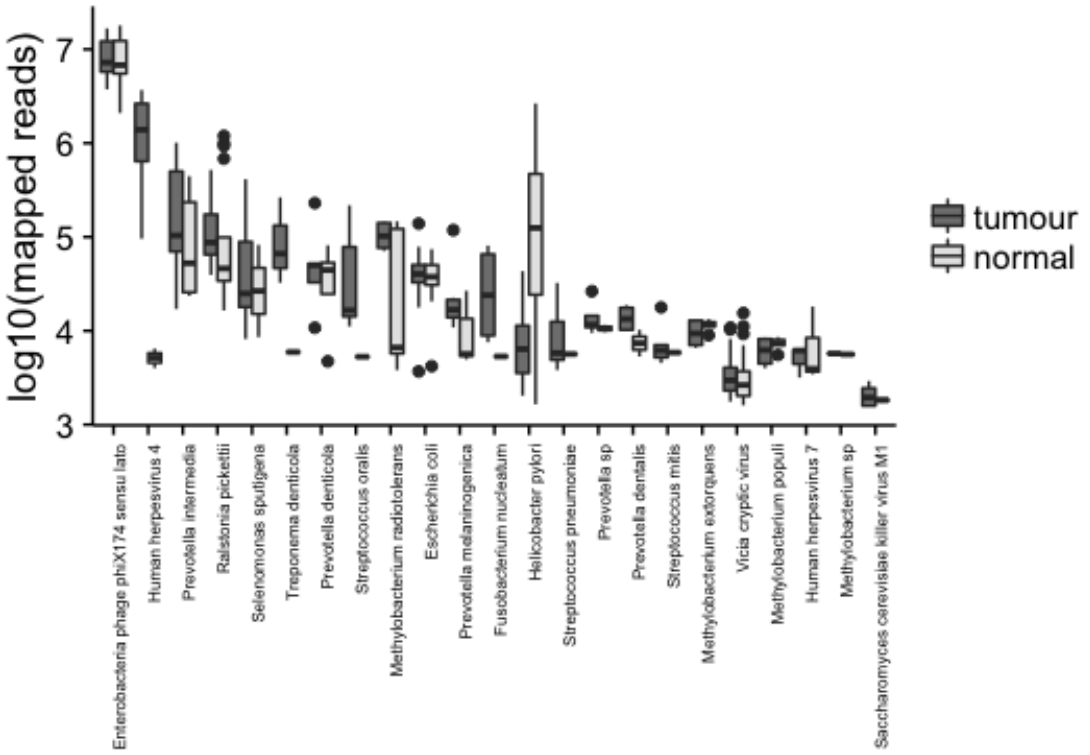


Figure 4.5.4: Boxplot illustrates the distribution of log10(mapped reads) for each species (x-axis) detected by the poshTitan in gastric cancer libraries. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

4.5.2 poshTitan reliably detects EBV and identifies *H. Pylori* in gastric cancer

As a published dataset, the cohort of 100 matched samples from patients with gastric cancer was screened for *EBV* with in-situ hybridization and for *H. Pylori* by inspection of either parallel gastric endoscopic biopsies or tumor resection. Thus, I wanted to explicitly compare detection as assayed by the poshTitan with experimental methods for both pathogens. I observed that poshTitan detected *EBV* in all samples known to be positive via in-situ hybridization and detected one additional positive, pfg125T (Table 4.1). Notably, the coverage of *EBV* in pfg125T was 97.75% (Figure 4.5.6), but interestingly this sample had the least number of reads mapped to the *EBV* reference genome by nearly two orders of magnitude (Figure 4.5.5). Thus, I suspect that poshTitan actually outperformed in-situ hybridizaton rather than detecting a false-positive. Including the *EBV* positive unique to the poshTitan analysis, poshTitan *EBV* detection was in agreement with parallel experimental methods in 99% of the 100 tumor samples.

Published in situ hybridization	**poshTitan** Detected	**poshTitan** Not Detected	**Total**
Detected	6	0	**6**
Not Detected	1	93	**94**
Total	**7**	**93**	**100**

Table 4.1: Table displays the number of samples in which EBV was detected in either the published in situ hybridization or by the poshTitan. If the in situ hybridization is treated as the ground truth, then the table displays the number of true positives (top left), failed detections (top right), false positives (bottom left), and true negatives (bottom right).

In contrast, detection of *H. Pylori* was much less consistent with experimental approaches (Table 4.2). poshTitan only detected *H. Pylori* in 29 of the 46 samples identified as positive by experimental methods including proximal normal samples

(Table 4.2). Interestingly, *H. Pylori* was detected in 3 patients that were identified as negative with experimental methods. *H. Pylori* is thought to provide an initiating event and does not necessarily colonize the tumor environment the way that *EBV* would as an intracellular viral pathogen. Moreover, I would expect bacteria to be subject to increased sampling effects compared to intra-cellular viruses. Thus, I was not expecting complete agreement between experimental methods and poshTitan. Despite this, I observed that 80 of 100 (or 80%) of poshTitan calls for *H. Pylori* were in agreement with published experimental annotation.

Presence in gastrectomy or gastric endoscopic biopsy	**poshTitan** Detected	Not Detected	**Total**
Detected	29	17	**46**
Not Detected	3	51	**54**
Total	**32**	**68**	**100**

Table 4.2: Table displays the number of samples in which *H. Pylori* was detected in either the published in situ hybridization or by the poshTitan. If presence in gastrectomy or gastric endoscopic biopsy preceding surgery is treated as the ground truth, then the table displays the number of true positives (top left), failed detections (top right), false positives (bottom left), and true negatives (bottom right).

4.5.3 Human herpesvirus 5 (*cytomegalovirus*) was identified in two patients' tumor samples

As previously discussed, I observed three human herpesviruses in the gastric cohort including *Human herpesvirus 4* (*EBV*). The other two herpesviruses observed were *Human herpesvirus 5* (*cytomegalovirus*), and *Human herpesvirus 7*. Notably, the patients in which these three herpesviruses were detected was mutually exclusive, i.e. no patient was positive for more than one of the three herpesviruses (Figures 4.5.5 and 4.5.6). While the detection of *Human herpesvirus 7* was intriguing and seemed robust with a high number of reads mapped (Figure 4.5.5) and

almost complete coverage (Figure 4.5.6), it was not consistently observed in tumor samples when observed in normal samples. This could be a result of tissue sampling or could suggest that the detection of *Human herpesvirus 7* is coincidental. In contrast, *cytomegalovirus* was exclusively detected in two patients' tumor samples with high coverage and number of mapped reads (Figures 4.5.5 and 4.5.6). Unfortunately, I was not able to obtain further tissue to validate this result with direct experimental methods such as in situ hybridization. However, given the discussion surrounding *cytomegalovirus* and cancer, these observations merit further screening and exploration. It is also important to note that *Human herpesvirus 7* and *cytomegalovirus* are in the same sub-family.

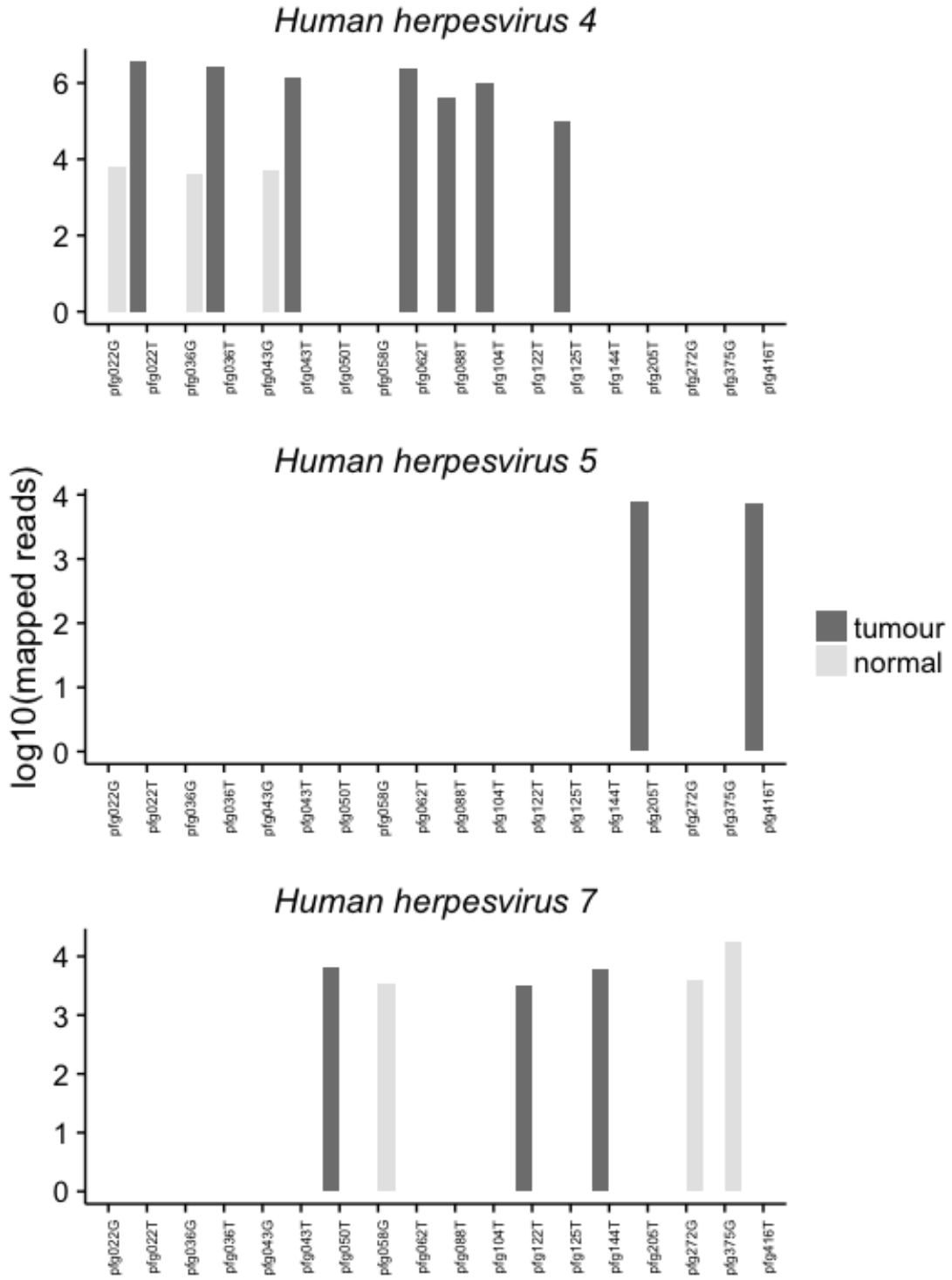


Figure 4.5.5: Barplot displays log2(mapped reads) (y-axs) of three human herpesviruses, Human herpesvirus 4 or EBV (top), human herpesvirus 5 or cytomegalovirus (middle), and human herpesvirus 7 (bottom) for each patient sample (x-axis) where the poshTitan detected a positive hit for any one of the herpesviruses in tumor (red) or normal (blue).

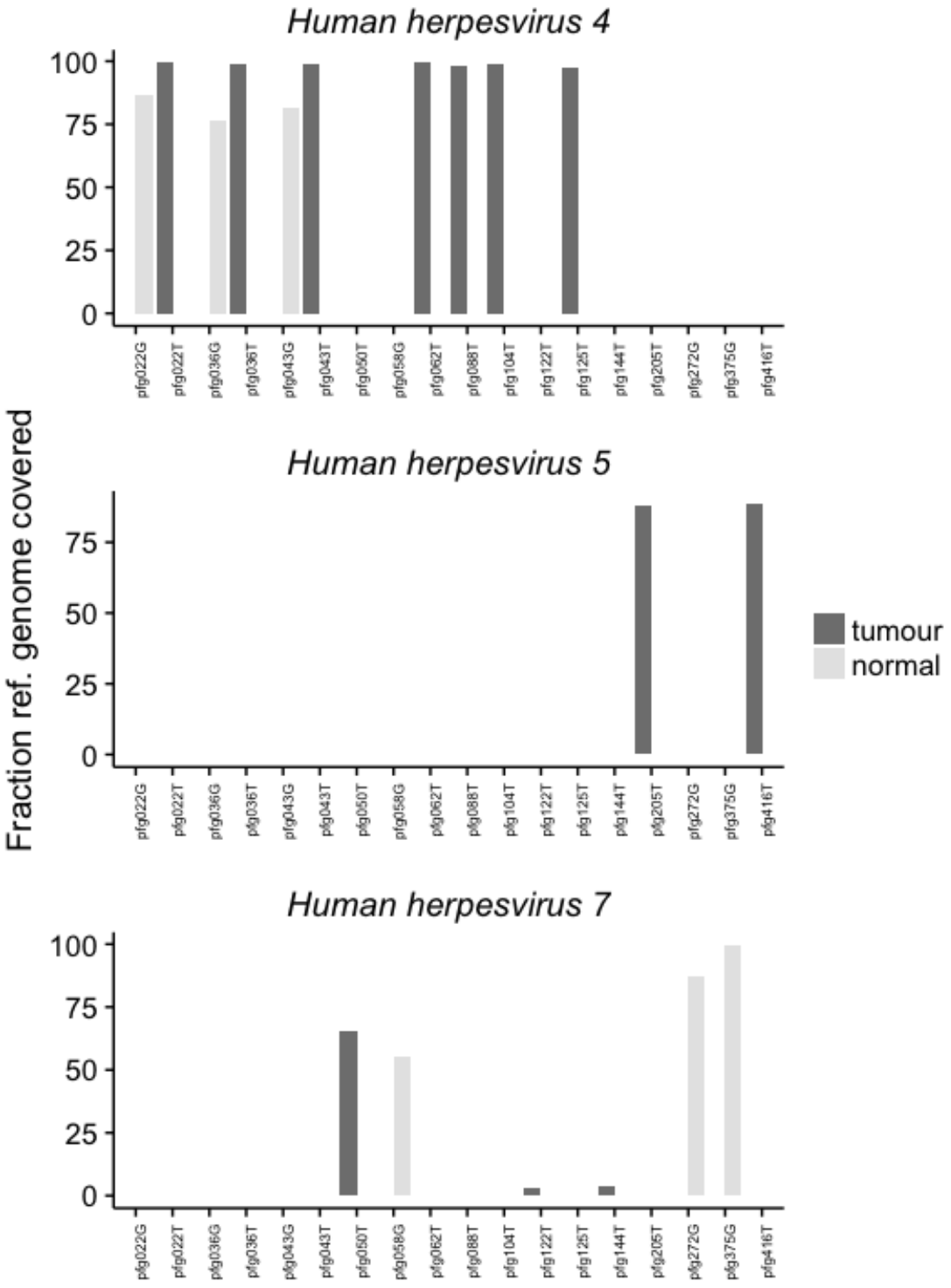


Figure 4.5.6: Barplot displays the fraction of the viral genome covered (y-axs) for three human herpesviruses, Human herpesvirus 4 or EBV (top), human herpesvirus 5 or cytomegalovirus (middle), and human herpesvirus 7 (bottom) for each patient sample (x-axis) where the poshTitan detected a positive hit for any one of the herpesviruses in tumor (red) or normal (blue).

4.6 poshTitan identifies EBV as ubiquitous in nasopharyngeal carcinoma

After investigating the performance of poshTitan on whole genome sequencing data, I wanted to understand how the algorithm performed with expression data. Through collaboration with Professor Ni, I obtained RNA-sequencing of matched nasopharyngeal carcinoma (NPC) tumor and proximal normal nasopharynx from 32 patients. NPC is known to be frequently associated with *Human herpesvirus 4* (*EBV*). As another positive control for the poshTitan, I wanted to determine if the algorithm could detect *EBV* in tumor and normal samples from the 32 patients. I observed that *EBV* was detected in every NPC tumor sample and none of the normal samples (Figure 4.6.1). This observation reflects the known association between *EBV* and NPC, and perhaps suggests that the association is even stronger than previously thought. Given the data resulted from transcriptome analysis, it was not surprising that coverage of *EBV* (Figure 4.6.2), while still well above the 2% threshold, was below that observed in the 7 *EBV* positive gastric tumors. *EBV* detected in tumors had the second highest number of reads mapping to the genome following *E. coli* (Figure 4.6.3). The number of hits observed was considerably less than in samples from patients with gastric cancer, which is not surprising since only expressed genes could be detected and holistic sequencing depth in RNA-seq is considerably less than whole genome sequencing. As in the gastric cohort, I observed several known contaminants, and, curiously, a subset of patients were positive for *Vicia cryptic virus* albeit with low coverage of its reference genome (Figures 4.6.1 and 4.6.2).

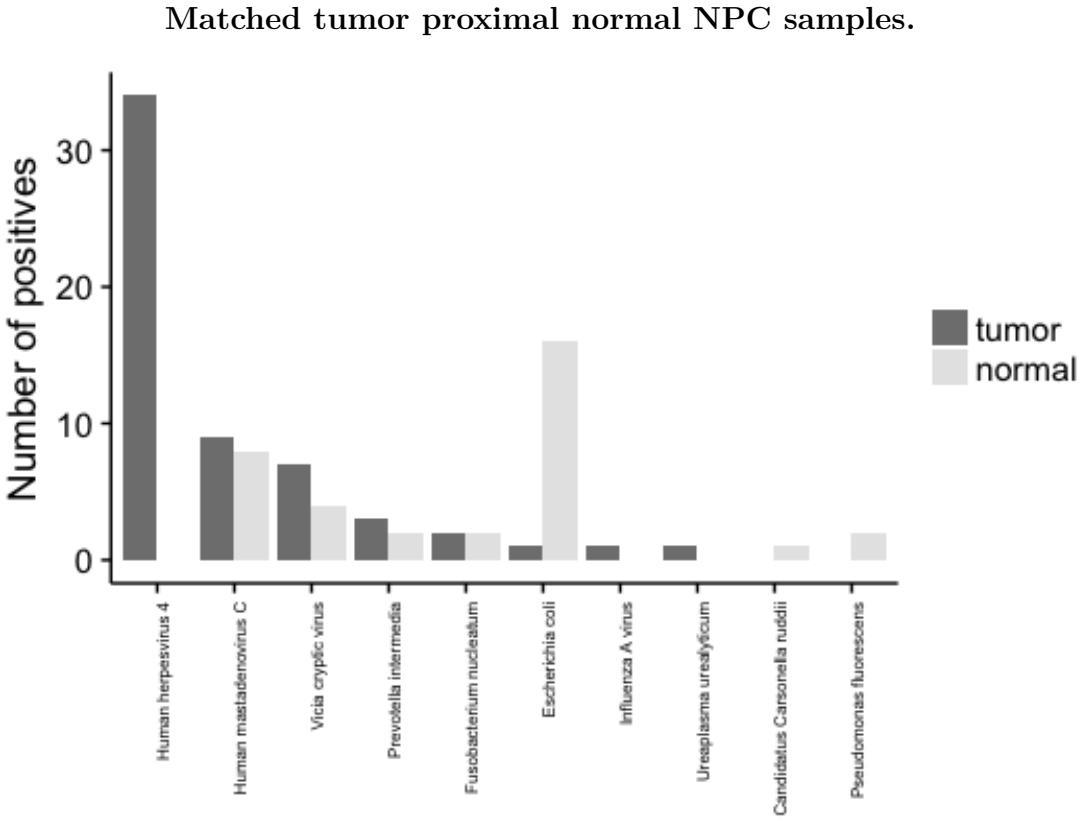


Figure 4.6.1: Barplot depicts the pathogens and commensals detected in both NPC tumors and proximal normal nasopharyngeal samples by poshTitan. The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) for tumor (dark grey) and normal (light grey) samples.

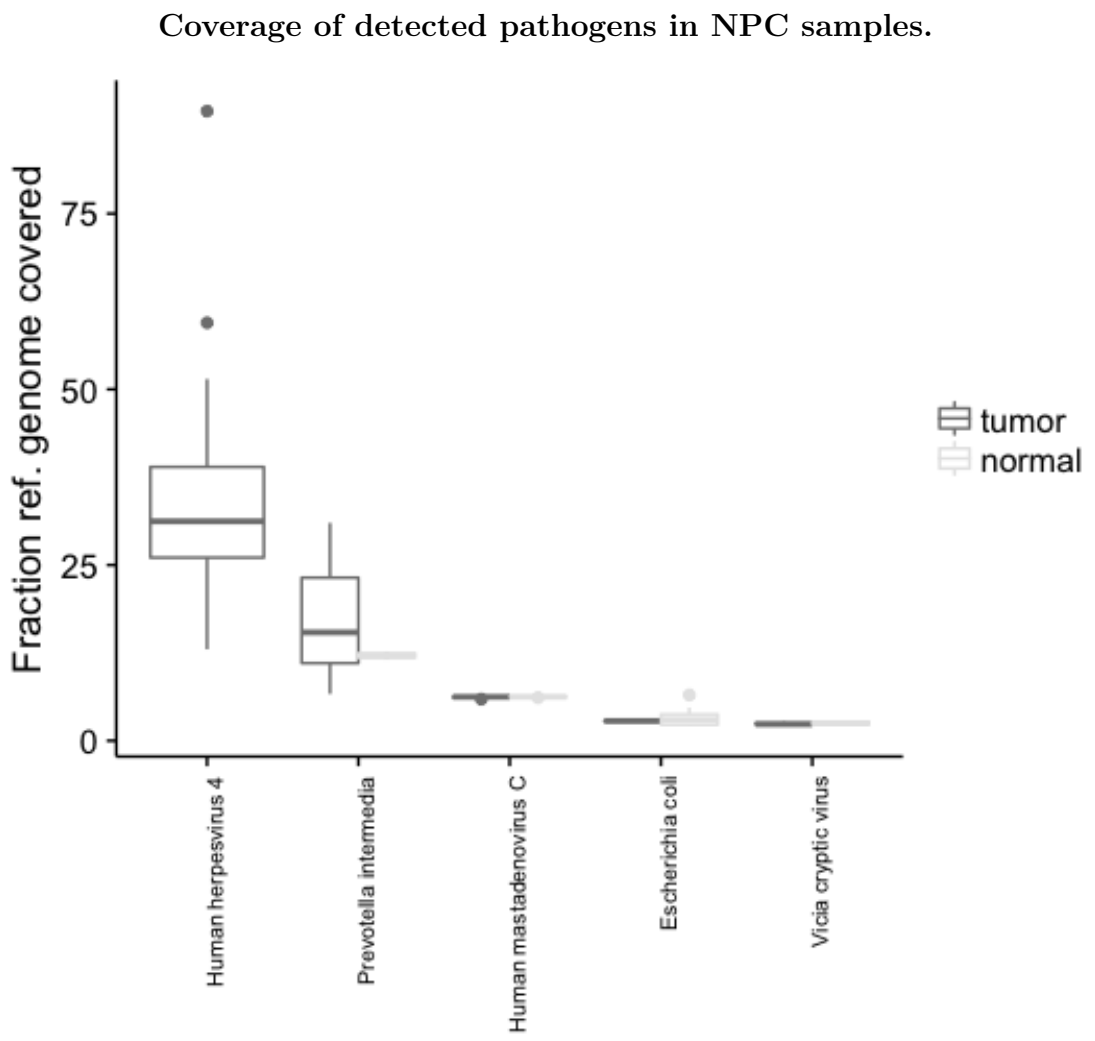


Figure 4.6.2: Boxplot illustrates the distribution of coverage for each species (x-axis) detected by the poshTitan in NPC libraries from tumor (dark grey) and proximal normal nasopharynx (light grey). The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

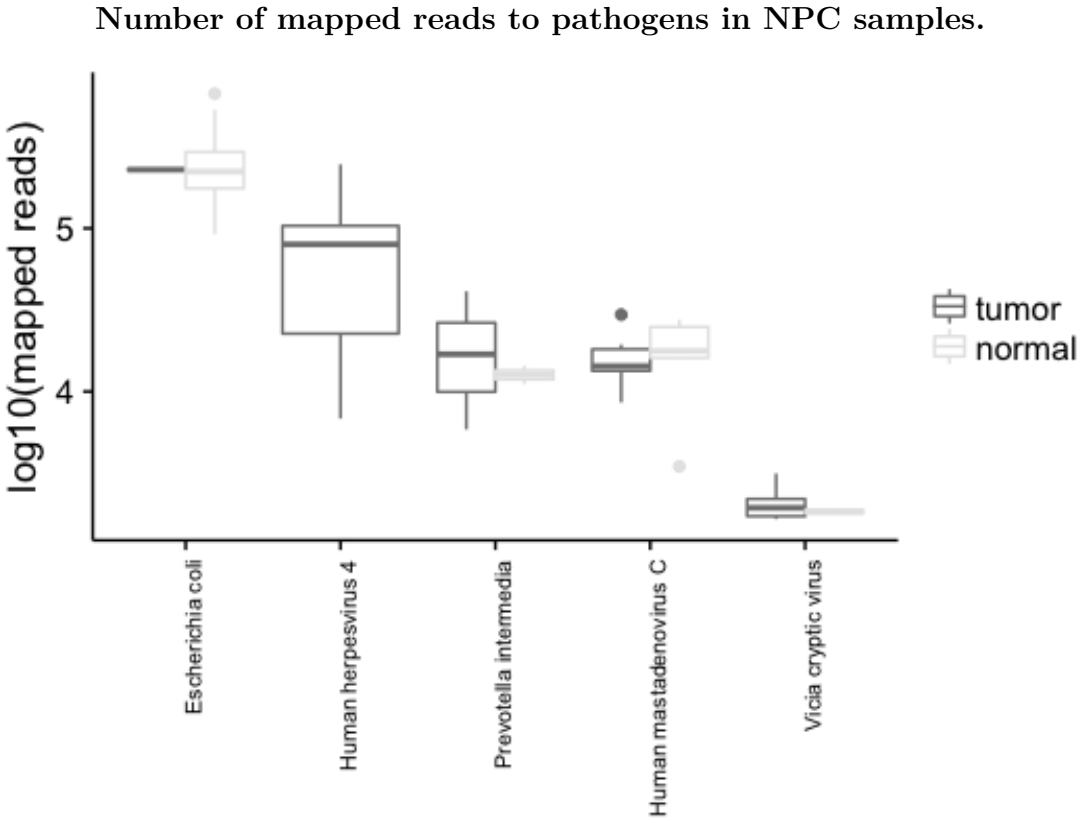


Figure 4.6.3: Boxplot illustrates the distribution of log10(mapped reads) for each species (x-axis) detected by the poshTitan in NPC libraries from tumor (dark grey) and proximal normal nasopharynx (light grey). The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

4.7 Pathogens do not appear to be important in Barrett's esophagus

After verifying that poshTitan could detect pathogens known to be associated with gastric cancer and nasopharyngeal squamous cell carcinoma in whole genome and RNA-sequencing experiments, respectively, I turned my attention to our own cohort of tissue sampled from patients that presented with Barrett's esophagus. Using poshTitan, I analyzed 16 whole tissue libraries taken from four patients, where tissue was taken from the stomach, duodenum, proximal normal esophagus, and the Barrett's lesion in each patient. From proximal tissue in the same biopsies, over 2000 single cells were isolated and sequenced as previously described (Chapter 4). These were also analyzed with poshTitan.

4.7.1 No pathogen association with Barrett's esophagus is observed in whole tissue libraries

In whole tissue libraries, poshTitan detected several pathogens and commensals across various tissue types (Figure 4.7.1). Most of the results belonged to microbes associated with commensalism (*Streptococcus*, *Lactobacillis*, *E. coli*) or contamination (*Rothia*, *Prevotella*, *Fusobacterium*). As in previous cohorts observed that bacteriophage was completely covered (Figure 4.7.2) and had the highest number of mapped reads (Figure 4.7.3). Similar to previously, the *Vicia cryptic virus* was detected in a subset of samples with extremely low coverage (Figures 4.7.1 and 4.7.2). After the third observation of this virus, I hypothesized that a portion of the *Vicia cryptic virus* may be shared wth a retrotransposon not included in the hg19 reference genome. However, this assertion is purely speculation and more work is necessary to substantiate or refute this hypothesis. On the balance, there were no striking associations with Barrett's tissue or any tissue in general in the poshTitan analysis of the whole tissue samples.

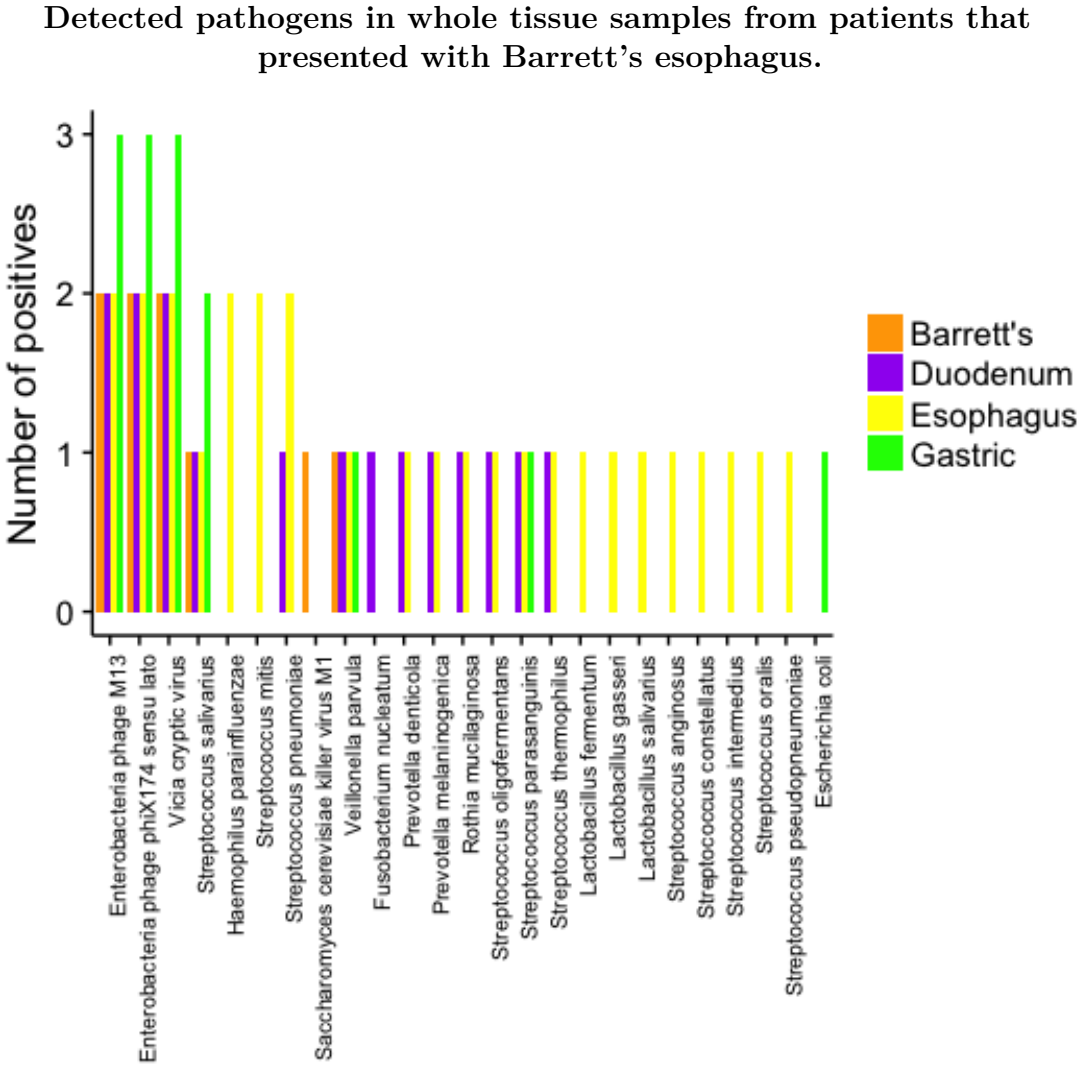


Figure 4.7.1: Barplot depicts the pathogens and commensals detected by the poshTitan in whole tissue libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) for tumor (dark grey) and normal (light grey) samples.

Coverage of detected pathogens in whole tissue samples from patients that presented with Barrett's esophagus.

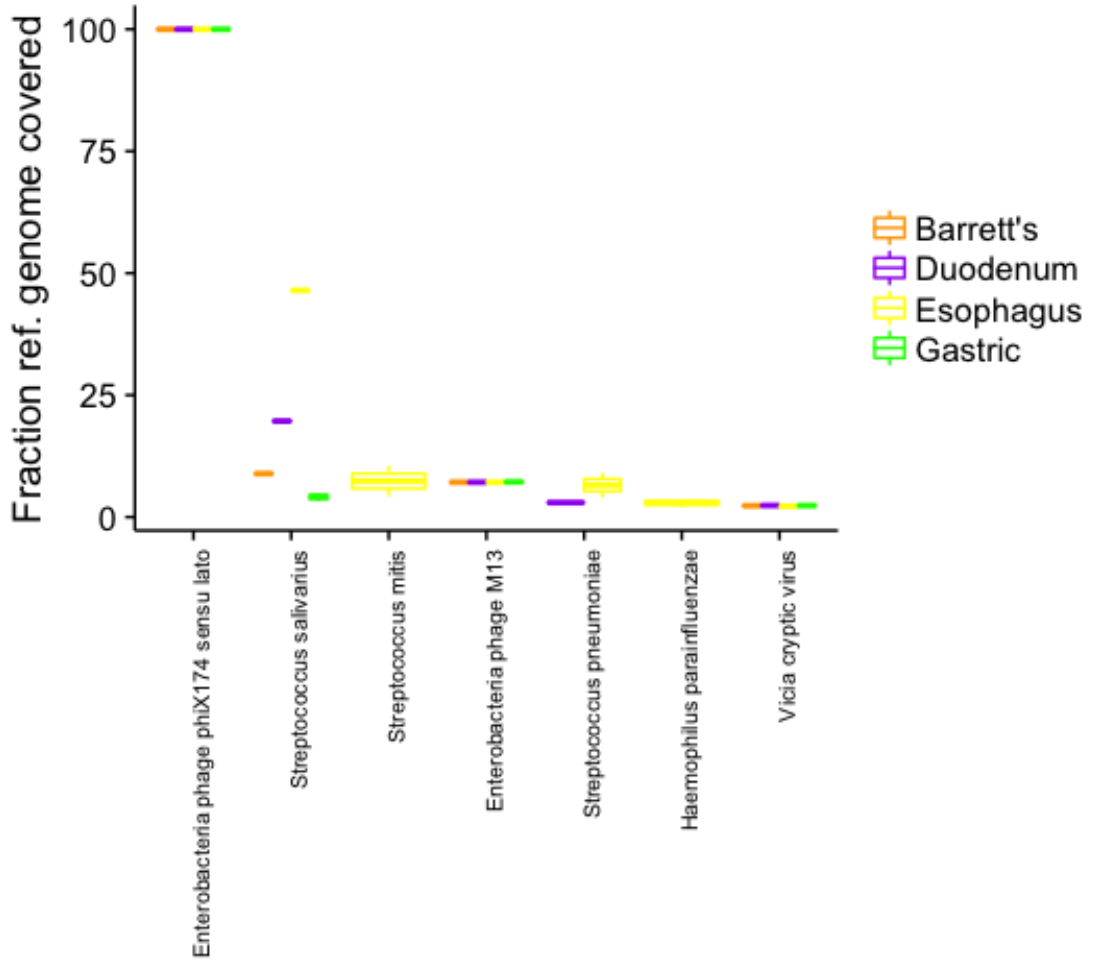


Figure 4.7.2: Boxplot illustrates the distribution of coverage for each species (x-axis) detected by the poshTitan in whole tissue libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

Number of mapped reads to pathogens in whole tissue samples from patients that presented with Barrett's esophagus.

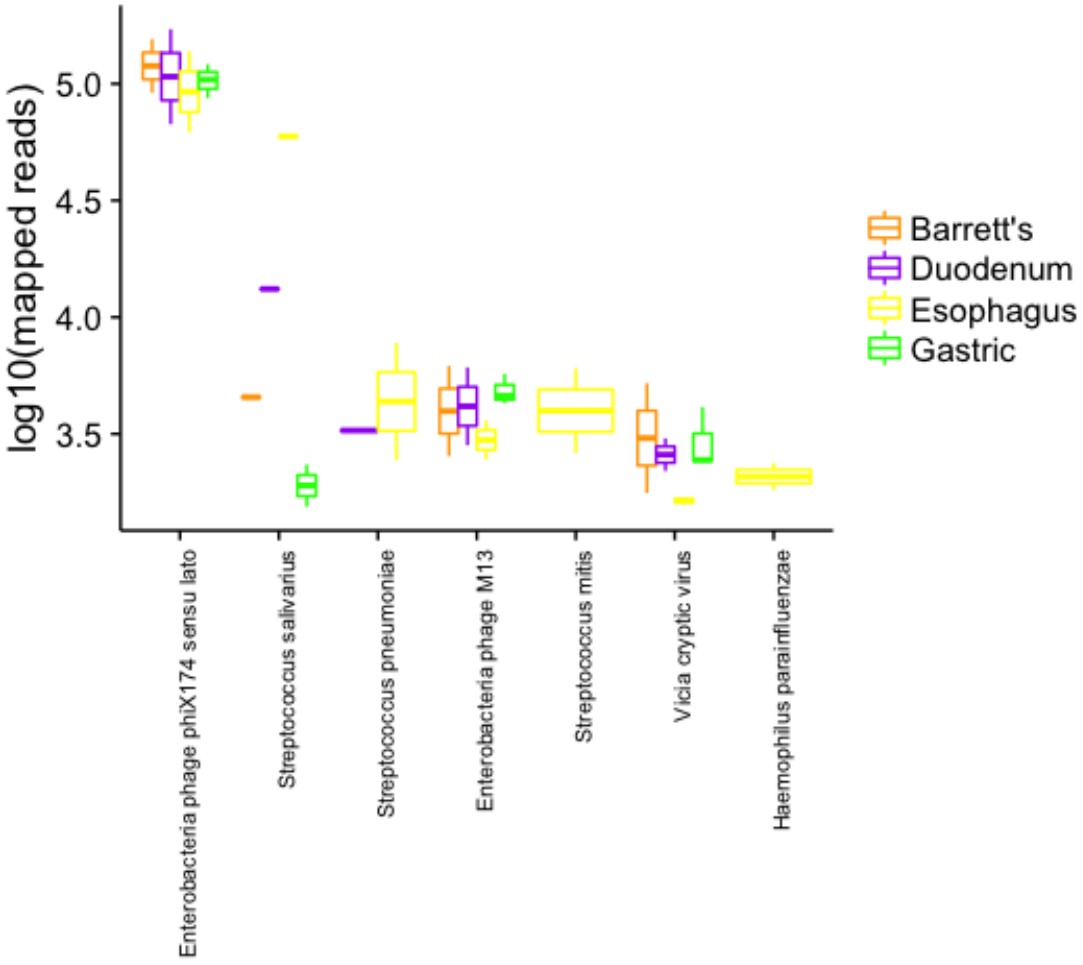


Figure 4.7.3: Boxplot illustrates the distribution of log10(mapped reads) for each species (x-axis) detected by the poshTitan in whole tissue libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

4.7.2 No pathogen association with Barrett's is observed in single cell libraries

After finding no striking association between any pathogens in whole tissue libraries, I resolved to examine the single cell libraries with the poshTitan, if only as a proof of principle. There were a number of pathogens only detected in 1 or 2 cells, but none of these were worthy of note, i.e. no observed pathogen had been previously characterized with some biological significance related to Barrett's esophagus or any cancer of the aero-digestive tract that would warrant further investigation despite only being observed in one or two cells (Supplementary Figure A.1.53). Amongst pathogens detected in at least three cells of any tissue type poshTitan detected two that have been published to likely be the result of contamination (*Micrococcus* and *Propionibacterium*) (Figure 4.7.4). In a handful of Barrett's and esophageal cells, poshTitan detected a pathogen identified in one of the whole tissue libraries and not detected in any of the 48 blank well libraries, *Veillonella parvula*. However, this pathogen is known to be a common dental and oral commensal, so its presence in the esophagus is not surprising. Similarly, a likely commensal *Streptococcus salivarius* was also detected in Barrett's and esophageal cells, but not in blank cells; however, it is also a well-characterized oral commensal. The most prevalently detected pathogen was *E. coli*, especially in cells derived from gastric tissue (Figure 4.7.4). But blank cells were also positive for *E. coli* and in blank samples the coverage of the *E. coli* was even greater than that observed in actual cell samples (Figure 4.7.5). This along with the high number of reads mapped to *E. coli* observed in blank cells (Figure 4.7.6), suggested that the frequency of *E. coli* positives in this data *E. coli* was likely a contamination artifact. Holistically, I determined that no interesting associations between pathogens and Barrett's esophagus were present in either the whole tissue or single cell RNA-sequencing libraries, i.e. no pathogens exhibited any tissue specific patter or were characterized by any known biological of obvious relevance to Barrett's esophagus or cancer more generally.

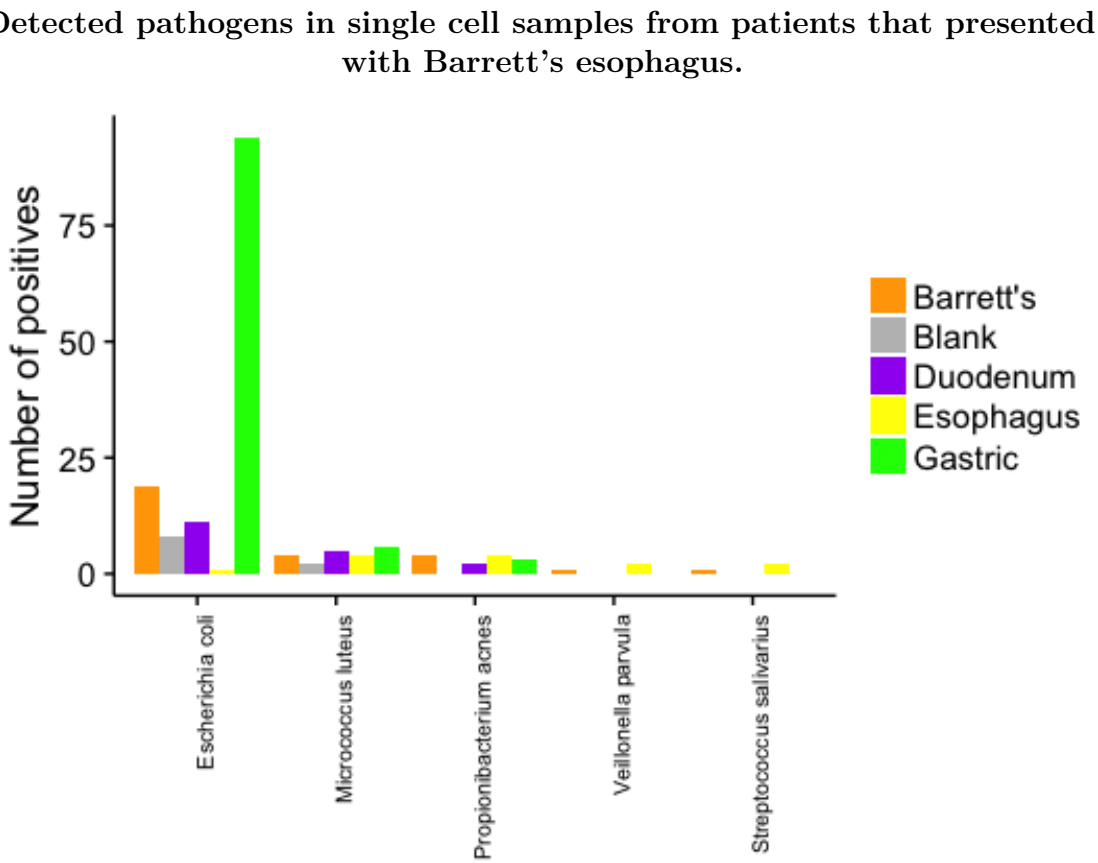


Figure 4.7.4: Barplot depicts the pathogens and commensals detected by the poshTitan in at least three single cell libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). Brain whole tissue positive control (none detected) and blank water negative control (grey) samples were also analyzed. The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) for tumor (dark grey) and normal (light grey) samples.

Coverage of detected pathogens in single cell samples from patients that presented with Barrett's esophagus.

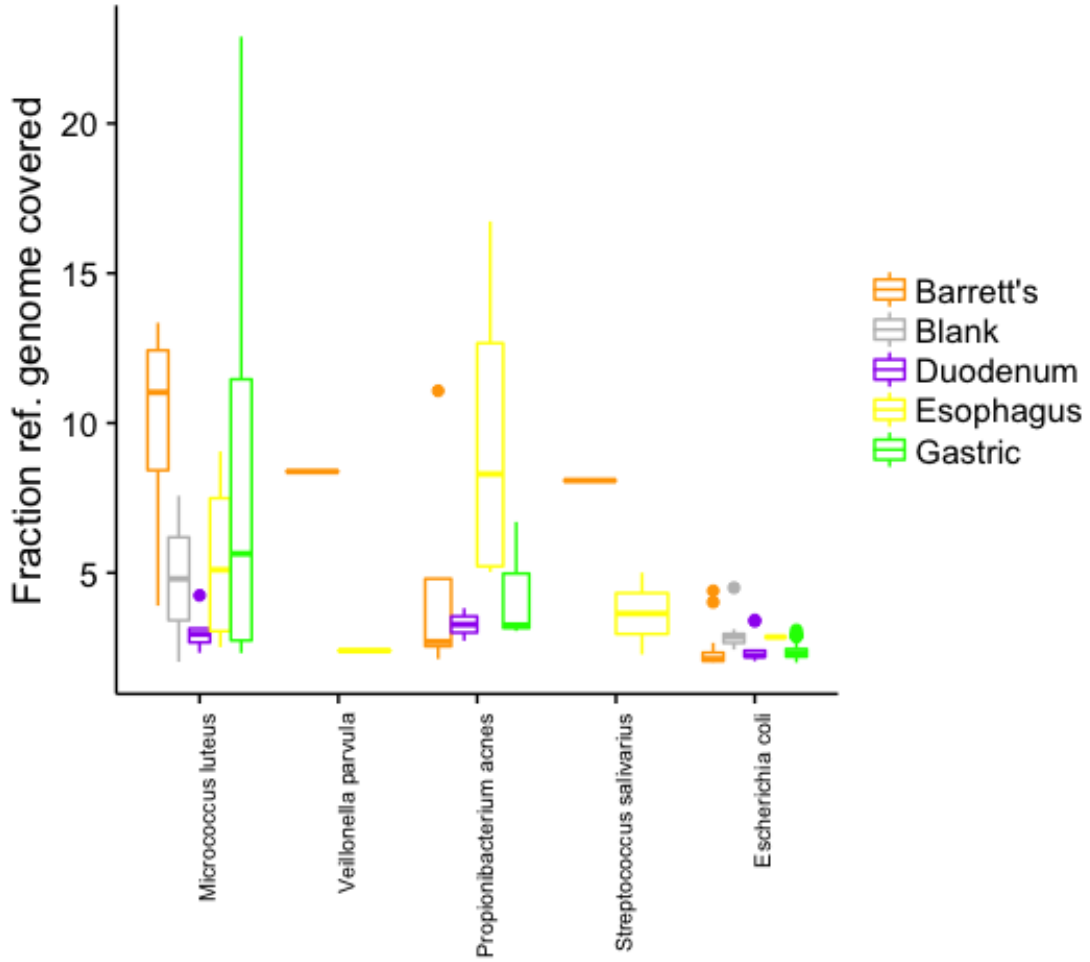


Figure 4.7.5: Boxplot illustrates the distribution of coverage for each species (x-axis) detected by the poshTitan in single cell libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). Brain whole tissue positive control (none detected) and blank water negative control (grey) samples were also analyzed. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

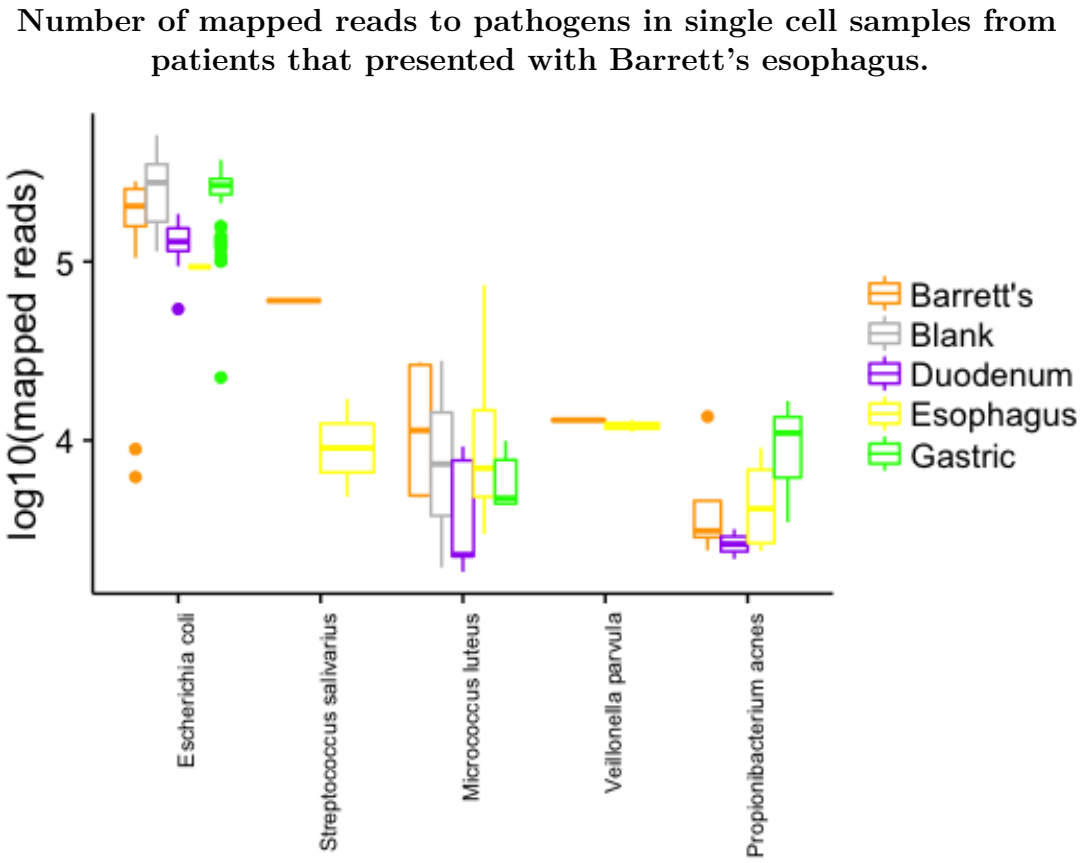


Figure 4.7.6: Boxplot illustrates the distribution of log10(mapped reads) for each species (x-axis) detected by the poshTitan in single cell libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). Brain whole tissue positive control (none detected) and blank water negative control (grey) samples were also analyzed. The bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots.

4.8 Conclusions and discussion

Here I have described the poshTitan, a pipeline for identifying pathogens and commensals in next-generation sequencing data. Although the poshTitan requires some tuning and further work before being distributed to a wider community, I was able to demonstrate that, in its primitive form, the algorithm could accurately identify pathogens present in simulated data with no false-positives (with adequate thresholds) and with sensitivity limited only by extremely low read numbers. Further work will aim to rigorously design simulations with the purpose of explicitly characterize the sensitivity and accuracy, and in order to acquire a more nuanced understanding of a coverage threshold for calling the presence of a pathogen. Indeed, coverage is not ideal as a heuristic for identifying the presence or absence of pathogens; some form of statistic should be developed based on the coverage conditional upon the total number of reads mapped to the pathogen reference, the size of the pathogen reference itself, and the sequencing depth of the library. Additionally, a great deal of work remains in terms of logistically packaging the poshTitan pipeline, so that it is portable to other researchers.

Nevertheless, poshTitan successfully identified *H. Pylori* in gastric cancer samples with 80% accuracy and *EBV* with 99% accuracy, potentially outperforming in situ hybridization. However, for the latter claim, further work includes performing a second in situ hybridization on remaining tumor sample in order to assess whether the first experiment was indeed a false positive. poshTitan was also able to demonstrate in 32 matched tumor and normal samples from patients with nasopharyngeal squamous cell carcinoma that *EBV* was unique to tumor samples and ubiquitously expressed. Finally, poshTitan provided no evidence that pathogens or commensals are important in Barrett's esophagus at the single cell or whole tissue level. This confirms the general consensus that Barrett's esophagus arises from environmental and molecular events intrinsic to the patient rather than from foreign pathogens or commensals.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe in *Wilhelm Meisters Wanderjahre, oder Die Entsagenden*

5 Conclusions and Discussion

In order to better understand potential intrinsic and extrinsic biology surrounding origins of the pre-malignant lesion Barrett's esophagus, I have developed two algorithms designed to enrich interpretations of next-generation sequencing experiments, and single cell RNA-sequencing in particular. The computational methods, BEARscc (Chapter 3) and poshTitan (Chapter 5) have been tested using control experiments. By sequencing dilute total RNA extracted from brain tissue and wells containing only water with spike-in molecules, I have shown that BEARscc accurately models technical variation. Furthermore, I have demonstrated that by simulating technical replicates, BEARscc can reconcile clustering artifacts that prominent sequencing algorithms tend to produce as a result of technical variation inherent to single cell RNA sequencing data. Benchmarking demonstrated that BEARscc generally performs well as a new clustering mechanism. Using data from two recently published single cell sequencing cohorts from the murine cortex and hippocampus and from the murine intestine, I demonstrated how BEARscc can enhance analysis in real single cell data. In the mouse cortex and hippocampus, BEARscc directly presents an interpretation of the data that previously required substantial manual curation. In the mouse intestine, the identification of cells and their groupings that were sensitive to noise suggested a new interpretation of two previously described stem-like populations. Using simulated libraries of varying

pathogen content and number of point mutations in reads, I demonstrated the ability of poshTitan to accurately identify pathogen reads in next-generation sequencing data; this experiment also illustrated how the pipeline provides an improvement over Kraken (Wood & Salzberg 2014) software alone. By examining data from collaborators from patients either presenting with gastric cancer or nasopharyngeal carcinoma I demonstrated the accuracy of poshTitan and its ability to identify known pathogens in whole genome and RNA sequencing data, respectively.

Still, there is certainly scope for optimization and improvement in both the BEARscc and poshTitan methods. It is necessary in future work to optimize its performance and more rigorously evaluate its sensitivity and propensity to detect false positives. A more thorough design of of simulation experiments aiming to would do much to improve our understanding of the limits of the algorithm. Specifically, an experiment to define the exact detection limit of poshTitan in terms of genomic diversity relative to the reference and the number of reads present in the genome would facilitate more accurate binary calls of pathogen content in data. In addition, ideally an *in vitro* experiment involving the sequencing of varied and titrated admixtures of pathogen DNA to the same depth could help define the precise threshold for excluding false positives while maximizing sensitivity. Together these experiments could be used to develop a more rigorous statistic than simple coverage of the reference genome for defining detection that coverage conditional upon the total number of reads mapped to the pathogen reference, the size of the pathogen reference itself, and the sequencing depth of the library.

Although BEARscc implementation and benchmarking are much more mature than poshTitan, there are a few obvious improvements for BEARscc going forward. Notably, while the drop-out model currently produces simulated replicates within 5% for the majority of genes, this distortion still meaningfully impacts genes with the least expression on average (less than 5 counts). From this empirical perspective and from a mathematical perspective, it would be ideal to adapt the general concept of the spike-in derived drop-out recovery and injections distributions and the single parameter noise model into a single coherent and rigorous bayesian framework. We

suspect that this would improve the simulated drop-out rate discrepancy, and as a corollary the divergence between simulated and observed gene expression for genes with both rare and low gene expression. In terms of benchmarking, there are two clear points of improvement. First, currently BEARscc has been benchmarked as a clustering algorithm against the naive clustering method and a simple gene sampling. Yet it would be better to identify a way to benchmark BEARscc against a simple gene sampling in terms of its ability to identify cells and clusters that are robust to noise as this is the primary function for which the algorithm is intended. In a similar vein, when we applied various clustering algorithms to the murine brain and intestine cohorts in order to benchmark BEARscc, we inevitably added an element of clustering method comparison because the ground truth in the datasets in question were defined by RaceID2 and BackSPIN, respectively. It would be better to create a dataset, perhaps from several cell lines, whereby the ground truth was pre-defined much like the brain and blanks experiment except with genuine single cells. This would provide a fairer method of benchmarking on all fronts, especially if an alternative benchmarking calculation were developed that could compare the sampling and noise consensus results to the ground truth directly.

In the two algorithms' current forms, we were able to glean new insights into the biology of Barrett's esophagus as a result of their application to whole tissue and single cell RNA sequencing libraries. Applying the poshTitan to whole tissue and single-cell libraries from upper gastro-intestinal endoscopic biopsies failed to identify any pathogens or commensals with any meaningful association with Barrett's esophagus, which supported the prevailing hypothesis that pathogens and commensals are not important in Barrett's esophagus. Biological observations illuminated by BEARscc were more extensive. Using robust subpopulations of cells derived from the BEARscc noise consensus, we were able to understand the functional subpopulation structure in three normal tissues of the human gastro-intestinal tract and the Barrett's lesion. From these insights within tissues, coupled with simultaneous clustering of all samples across tissue types, we were able to observe that esophageal submucosal glandular ducts and undifferentiated Barrett's

esophagus were functionally similar and that similarity was invariant to experimental noise as modeled by BEARscc. Furthermore, differential expression analysis of robust cells within clusters identified LEFTY1 and stem cell associated OLFM4 as characteristic of this subpopulation along with two new markers for early goblet cells relevant to the colon, Barrett's esophagus, and esophageal submucosal glandular ducts. The identification of these marker genes was verified at the protein level by Dr. Michael White.

There are numerous lines of further investigation following these new findings in Barrett's esophagus. Among the most pressing is a validation experiment whereby pathologically verified esophageal submucosal glandular ducts in patients without Barrett's esophagus are micro-dissected and processed into a single cell suspension as done with the samples analyzed in this work. Albeit logistically challenging due to the infrequency of esophageal submucosal glandular ducts in the normal esophagus, this experiment would do much to conclusively refute or support the results as they observed and interpreted in this work. Following the success of the Hans Clever lab with LGR5 positive cells in the intestine (Barker *et al.* 2007), a lineage tracing experiment in small intestine organoids, whereby ITLN1 and/or SPINK4 are genetically modified to express EGFP or some other fluorescent protein would provide informative data verifying the early marker gene hypothesis presented here as a result of computational and *in vivo* immunofluorescence observations. A similar approach could be taken with OLMF4, although ideally this experiment would be performed in organoid models of esophageal submucosal glandular ducts and Barrett's, but as of yet neither of these models exist.

Collectively, this work presents important contributions in terms of both the biology of Barrett's esophagus and bioinformatic methods in next-generation sequencing analysis. Most importantly, these data and methods raise new questions within the scope of the relevant fields as any decent science should. A number of questions arise surrounding the possibility of Barrett's originating from esophageal submucosal glandular ducts. Assuming future work corroborates our results, and

esophageal submucosal glandular ducts are the origin of Barrett's esophagus, then new questions should be pursued. Among them:

- Do all presentations of Barrett's esophagus arise from esophageal submucosal glandular ducts, or are there multiple routes to Barrett's esophagus?
- Could malignancy arise in esophageal submucosal glandular ducts before the putative glandular takeover forms a Barrett's phenotype? Could this explain the low number of patients with Barrett's diagnoses prior to presentation with esophageal adenocarcinoma?
- Are there biomarkers that could be used to identify pre-malignancy in esophageal submucosal glandular ducts before a Barrett's phenotype even arises? Could these biomarkers be used to separate glandular ducts destined for stable pre-malignancy and those that will progress to esophageal adenocarcinoma?

Likewise, the analysis of control single cell RNA sequencing experiments using BEARscc in conjunction with various cluster algorithms raises several concerns, chief among them:

- Could we design a mechanism for identifying batch effects purely from the structure of the data in single cell RNA sequencing experiments?
- If so, how do we incorporate such a method into analytical tools such as BEARscc and is it possible to transform data plagued by batch effects to remove these distortions without distorting the real biological signal? Indeed, is this necessary in order to answer the pertinent questions in single cell biology that single cell RNA sequencing technology is expected to inform?

There are many challenges, and thus, exciting new directions as experimental and computational approaches continue to push forward our knowledge of Barrett's esophagus, esophageal adenocarcinoma, and single cell RNA sequencing analysis.

We shape our buildings; thereafter they shape us.

— Winston Churchill

6 Afterword

Contents

6.1 Beyond the scope of this work **211**
6.2 Relevant publications **215**

6.1 Beyond the scope of this work

The previous work describes methods in single cell RNA sequencing data analyses and their application to Barrett's esophagus. While the aforementioned work represents the primary project occupying the final 18 months of my time as a doctoral student at the University of Oxford, for reasons of scope and coherence this thesis does not address a number of research questions I pursued early in my degree as well as projects engaged in parallel with my work on Barrett's esophagus. One such work was a minor programmatic contribution enabling the identification and subsequent comparison of proteins with sequences predicting canonical nuclear localization signals and ankyrin repeats featured in my colleague Min Lu's publication in *Cell*. Of more significant note are three other projects that occupied the majority of my attention in the early years of my D.Phil. In broad terms, the three projects were as follows: identifying co-regulation of gene expression

by pairs of transcription factors *in silico* with a specific emphasis on p53 and NFκB; understanding the role of ASPP2 in neuro-inflammation, and finally examining the changes in host and viral gene expression as a function of radiotherapy response and viral genotype in nasopharyngeal carcinoma.

In the exploratory phase of my research degree, I initially pursued work that aimed to understand the role of transcription factors in gene regulation. Specifically, I aimed to understand how co-regulation of transcription factors led to changes in genome-wide gene expression in inflammatory and cancer contexts. In the first year of work, I spent some time attempting wet lab work, especially chromatin immunoprecipitation (ChIP); however, my chief intellectual advancements in that year occurred *in silico*. By designing a pipeline to associate pairings of transcription factor binding data obtained from ChIP sequencing data with relevant expression data from either RNA-sequencing or microarray platforms, I familiarized myself with the fields of genomics, transcriptomics, epigenomics and the relevant coding skills in perl, bash, and R pertaining to these subjects. Unfortunately, the finished pipeline and results identifying a handful of genes tying together p53 and NFκB will likely never be published due to time constraints and more interesting work that drew my attention. Despite this, I feel the exploration of this project during my first 18 months of doctoral work was essential in the initial phases of my graduate education. Having arrived at the University of Oxford with a liberal arts and sciences degree concentrated in mathematics and biochemistry, this project was a catalyst for the development of computational skills and genomic knowledge required to be a successful computational biologist.

By developing these skills, I was in a unique position in our molecular cell biology lab to assist in a project lead by Dr. Casmir Turnquist aiming to understand the role of ASPP2 in neuro-inflammation. Dr. Turnquist initially pursued the project in order to understand a severe encephalopathic phenotype observed in ASPP2 knockout mice. In the course of her work, she discovered that both murine (macrophage, RAW264.7; microglial, BV-2) and and human (monocyte, THP-1) monocyte lineage cell lines would express ASPP2 protein in response to

lipopolysaccharide (LPS). Although NFκB activation was present in these cells, evidence mounted that the ASPP2 upregulation in response to inflammation was, in fact, NFκB dependent. Our supervisor, Professor Lu, then suggested we look at a number of transcription factors including STAT1, and this is where I began to be involved in the project. First, we hoped to corroborate Dr. Turnquist's findings in the wet lab by mining cell line and inflammatory disease microarray data. To that end, Dr. Turnquist curated a list of relevant studies involved in neurodegenerative, myeloid cell inflammatory stimulus treatment, and inflammatory disease. I then systematically re-analyzed the microarray expression data of these studies and reported the fold change in ASPP2 expression in each case. The data unanimously supported the finding that inflammatory stimulus was associated with increased ASPP2 expression. Thus, the question of the project narrowed on understanding how ASPP2 expression was being induced following inflammatory stimulation (operationalized by LPS in these experiments).

In order to answer this question, I examined publicly available ChIP-sequencing datasets where the epitope targeted in the immuno-precipitation belonged to inflammatory signaling transcription factors. I observed that, in contrast to NFκB, AP-1, and IRF-3, ChIP sequencing data existed that supported STAT-1 binding in the 1kb region upstream of the ASPP2 promoter in the human cells. Furthermore, using regex pattern matching I affirmed the presence of multiple STAT-1 consensus sequences within the human genomic region upstream of the ASPP2 promoter in question. This finding was reinforced when I used genome paired alignment software to align the mouse and human 3kb region upstream of the ASPP2 promoter, and discovered that several STAT-1 consensus sequences were conserved in the mouse (albeit with a varied distance to the promoter). I also confirmed that the binding region upstream of ASPP2 contained pertinent histone marks for enhancer activity (interestingly, these marks were restricted to leukocyte and astrocyte cell lines or primary tissue); this further supported the hypothetical presence of a STAT1 activated enhancer. To define the enhancer more precisely and verify it as a bonafide STAT-1 target site, I designed primers for a STAT1-ChIP experiment that

would triangulate and confirm the location of STAT1 binding in a specific region of the sequence upstream of the ASPP2 promoter. Dr. Yihua Wang performed the ChIP experiment and the putative enhancer was validated in both the human and the mouse. Finally, in order to confirm that the STAT-1 binding was functional, these primers were utilized to clone luciferase vectors with DNA upstream of the ASPP2 promoter. Dr. Wang cloned one vector containing the human enhancer and a negative control with no enhancer (though both were DNA from upstream of the ASPP2 transcription start site). The final luciferase assay performed by Dr. Turnquist in 293T cells showed increased activation of luciferase expression upon STAT1 activation compared to the control. This confirmed the functional activity of the enhancer *in vitro*, and suggested that we had discovered a novel STAT1 dependent enhancer for ASPP2 that operated in myeloid lineage cells of mice and humans. These finding were published in *Proceedings of the National Academy of Sciences*, where Drs. Wang and Turnquist were co-first authors.

As the STAT1 regulation of ASPP2 project concluded toward the end of my second summer, I began to develop the poshTitan pathway discussed in Chapter 4. Around the same time, Professor Lu decided to involve me in a project examining the role of Epstein-barr virus (EBV) in nasopharyngeal carcinoma. Specifically, Professor Lu had recently received 18 paired tumor normal total RNA-sequencing libraries, and suggested that I work to analyze these libraries in the context of radiotherapy and viral-host interactions. Because the samples consisted of total RNA libraries prepared using a ribo-depletion protocol, the likelihood of observing viral expression simultaneous with host expression seemed high. Therefore, I designed a pipeline that mapped and counted expression for both host and viral genomes. Because the annotation cultures of the viral genome field and human genome filed are vastly different, where many useful tools are only compatible with human reference genome conventions, it was necessary to modify the EBV reference annotation. Therefore, I rigorously re-annotated the EBV genome to be compatible with tools such as STAR. Following this, a colleague, Felix Zhou, recently designed

a method for extracting an approximation of viral genotype from processed RNA-sequencing variant data using machine learning techniques. The method revealed four distinct strain types of EBV present in the tumor samples. The cohort for this project was recently completed, and as the BEARscc and Barrett's work are concluded, it is my hope to return to this work. To conclude this project, I need to perform differential expression analysis on the complete cohort nasopharyngeal carcinoma cohort in order to identify expression signatures in the host associated with viral strain and radio-sensitivity phenotype.

6.2 Relevant publications

It should be noted that the work described in this thesis has lead to the production of a number of manuscripts. Specifically, the BEARscc method for single cell cluster analysis described in Chapter 2 has resulted in a methods manuscript recently reviewed for publication in *Nature Communications*. The reviews were positive in nature and the revised manuscript should be published there shortly. The anticipated publication reference is as follows:

> **David T. Severson[1]**, Richard P. Owen[1,2], Michael J. White[1,2], Xin Lu[1]*, and Benjamin Schuster-Böckler[1]*. **BEARscc determines robustness of single-cell clusters using simulated technical replicates.** *Nat. Commun.* In Revision.
>
> [1]Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [2] Oxford University Hospital NHS Trust, John Radcliffe Hospital, Oxford, UK.
> *Corresponding authors

The quality control and biological work discussed in Chapter 3 has also recently been reviewed for publication in *Nature Communications*, and the requested revisions are being completed by my colleagues, Drs. Richard Owen and Michael White. The anticipated publication reference is as follows:

> Richard P. Owen[1]*, Michael J. White[1]*, **David T. Severson[1]***, Barbara Braden Braden[2], Adam Bailey[2], Robert Goldin[3], Lai Mun Wang[4], Nicholas D. Maynard[5], Angie Green[6], Paolo Piazza[6,7], David Buck[6], Chris P. Ponting[8], Benjamin Schuster-Böckler[1^] and Xin Lu[1^]. **Single**

cell transcriptomes reveal a likely origin of Barrett's esophagus. *Nat. Commun.* In Revision.

[1]Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [2]Translational Gastroenterology Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [3]Centre for Pathology, St. Mary's Hospital, Imperial College, London, UK. [4]Department of Pathology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [5]Department of Upper GI Surgery, Oxford University Hospitals, Oxford, UK. [6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [7]Faculty of Medicine, Department of Medicine, Imperial College London, London, UK. [8]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Crewe Road, Edinburgh, UK.
*These authors contributed equally to this work
^Corresponding authors

Additionally, the work described in the previous section of this afterword concerning the discovery of a STAT1 responsive enhancer regulating the transcription of ASPP2 in mice and humans was published in *Proceedings of the National Academy of Sciences* under the following reference:

Casmir Turnquist[a,1,2], Yihua Wang[a,1], **David T. Severson[a]**, Shan Zhong[a], Bin Sun[b], Jingyi Ma[a], Stefan N. Constantinescu[c], Olaf Ansorge[d], Helen B Stolp[b], Zoltan Molnár[b], Francis G. Szele[b], and Xin Lu[a,3]. **STAT1-induced ASPP2 transcription identifies a link between neuroinflammation, cell polarity, and tumor suppression.** *Proc. Natl. Acad. Sci.* April 30, 2014; 111: 9834-9839.

[a]Ludwig Institute for Cancer Research, Nuffield Department of Medicine [b]Departments of Physiology, Anatomy and Genetics [c]Ludwig Institute for Cancer Research and Université Catholique de Louvain, de Duve Institute, 1200 Brussels, Belgium [d] Neuropathology, University of Oxford, Oxford OX3 7DQ, United Kingdom.
[1]These authors contributed equally to this work.
[2]Present address: Laboratory of Human Carcinogenesis, National Cancer Institute, National Institutes of Health, Bathesda, MD 20892.
[3]To whom correspondence should be addressed.

I made a small scripting contribution designed to predict proteins with nuclear localization signals computationally and this resulted middle authorship of a publication concerning a novel nuclear transport sequence in *Cell*:

Min Lu[1], Jaroslav Zak[1], Shuo Chen[1], Luis Sanchez-Pulido[2], **David T. Severson[1]**, Jane Endicott[3], Chris P. Ponting[2], Christopher J. Schofield[4], and Xin Lu[1]*. **A Code for RanGDP Binding in Ankyrin Repeats Defines a Nuclear Import Pathway.** *Nat. Commun.* April 30, 2014; 111: 9834-9839.

[1]Ludwig Institute for Cancer Research, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7DQ, UK. [2]MRC Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, OX1 3PT, UK. [3]Northern Institute for Cancer Research, University of Newcastle, Newcastle, NE2 4HH, UK [4]Chemistry Research Laboratory, Department of Chemistry, University of Oxford, Oxford, OX1 3TA, UK.

*Correspondence

Finally, the extensive work discussed briefly in the afterword concerning the annotation of the Epstein-Barr virus reference genome in conjunction with the virus' role in nasopharyngeal carcinoma is still in progress, and a manuscript is expected to be forthcoming in one or two years.

Appendices

The only thing that will redeem mankind is cooperation.

— Bertrand Russel in *Human Society in Ethics and Politics*

A

Supplementary Figures, Tables, and Software

Contents

A.1 Supplementary Figures **222**
A.1.1 BEARscc benchmarking 222
A.1.2 Single cell cohort description 225
A.1.3 Esophageal cluster interpretation 230
A.1.4 Gastric cluster interpretation 238
A.1.5 Duodenal cluster interpretation 246
A.1.6 Barrett's tissue cluster interpretation 256
A.1.7 All QC-filtered cell cluster interpretation 266
A.1.8 Further analysis of all QC-filtered cells 281
A.1.9 Single cell pathogen detection 285
A.2 Supplementary Tables **287**
A.2.1 Single cell Barrett's cohort descriptions 287
A.2.2 BEARscc metrics and differential expression exclusion . . 288
A.2.3 Differentially expressed genes across the esophageal clusters 291
A.2.4 Differentially expressed genes across the gastric clusters . 315
A.2.5 Differentially expressed genes across the duodenal clusters 320
A.2.6 Differentially expressed genes across Barrett's tissue clusters 328
A.2.7 Differentially expressed genes in clusters from all QC-filtered cells 334
A.2.8 Differential expression summary 353
A.2.9 BEARscc metrics for clusters from all QC-filtered cells . 353
A.3 List of Software **354**

The following figures and tables include work that is important to the scope of

this thesis and thus referenced in the text, but was largely or entirely performed by others or was too cumbersome for inclusion in the main text. In the former case, the caption of the figure will attribute credit to whomever generated the data.

A.1 Supplementary Figures

A.1.1 BEARscc benchmarking

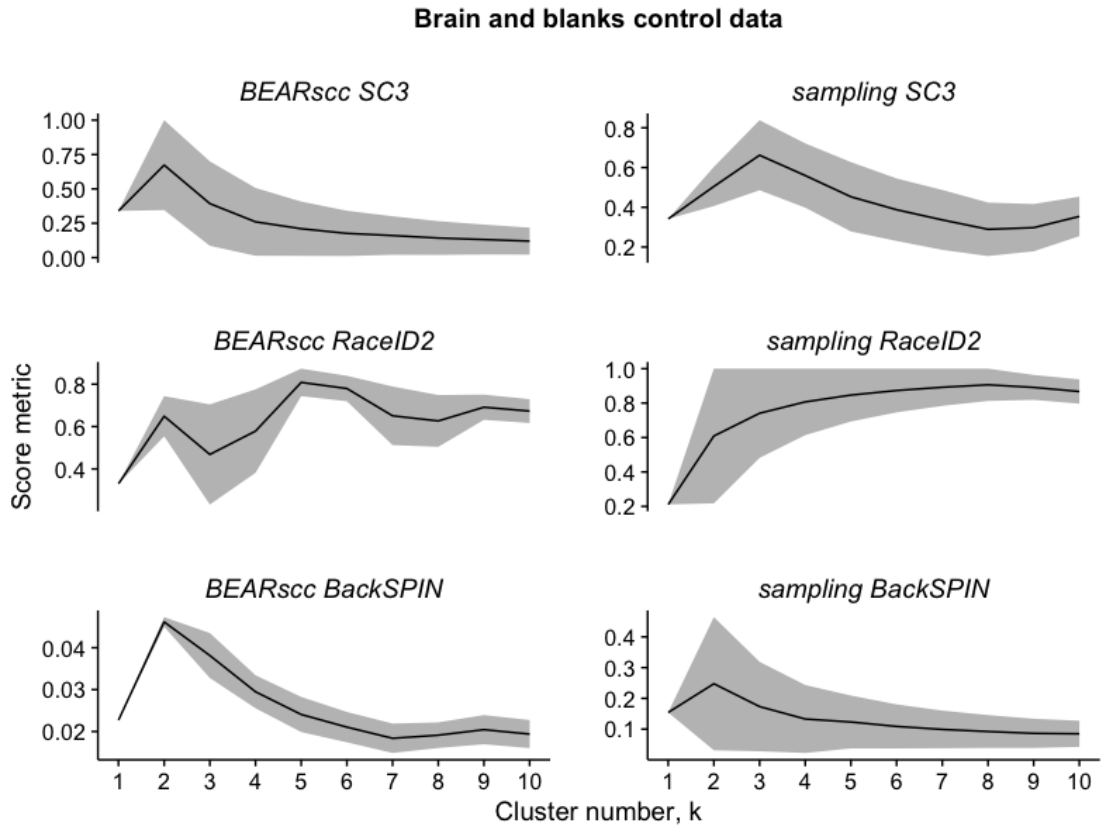


Figure A.1.1: The ribbon plot displays the mean BEARscc cluster score observed for each splitting of the dendrogram formed from hierarchical clustering of the noise or subsampling consensus matrix of the brain and blanks data. Mean value is indicated by the black line, while the grey area denotes standard error as indication of cluster score spread for each clustering.

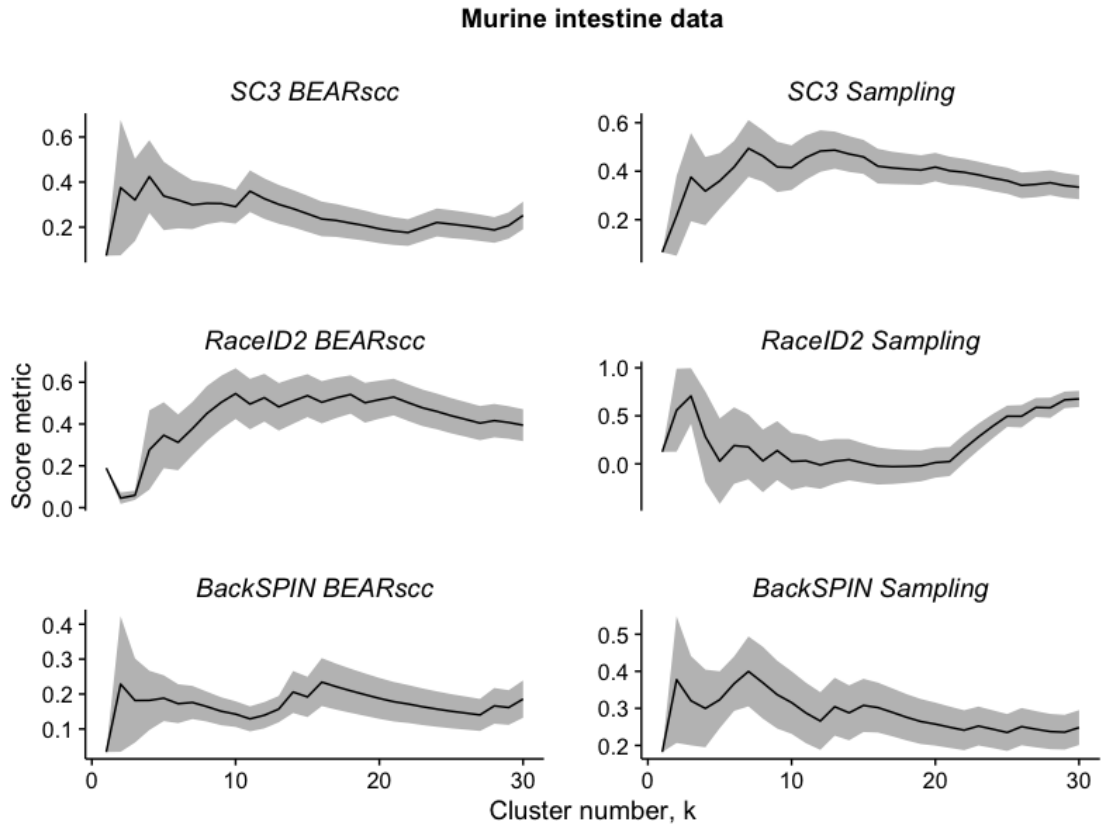


Figure A.1.2: The ribbon plot displays the mean BEARscc cluster score observed for each splitting of the dendrogram formed from hierarchical clustering of the noise or subsampling consensus matrix of the murine intestine data. Mean value is indicated by the black line, while the grey area denotes standard error as indication of cluster score spread for each clustering.

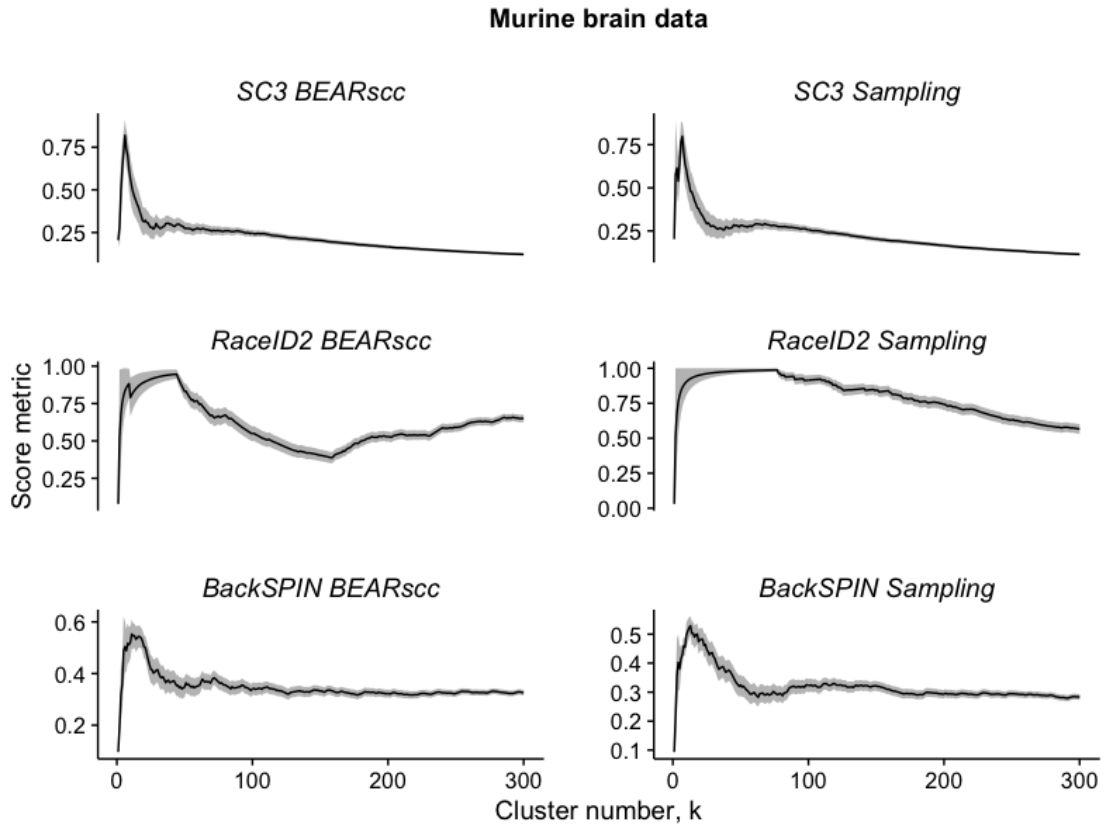


Figure A.1.3: The ribbon plot displays the mean BEARscc cluster score observed for each splitting of the dendrogram formed from hierarchical clustering of the noise or subsampling consensus matrix of the murine brain data. Mean value is indicated by the black line, while the grey area denotes standard error as indication of cluster score spread for each clustering.

A.1.2 Single cell cohort description

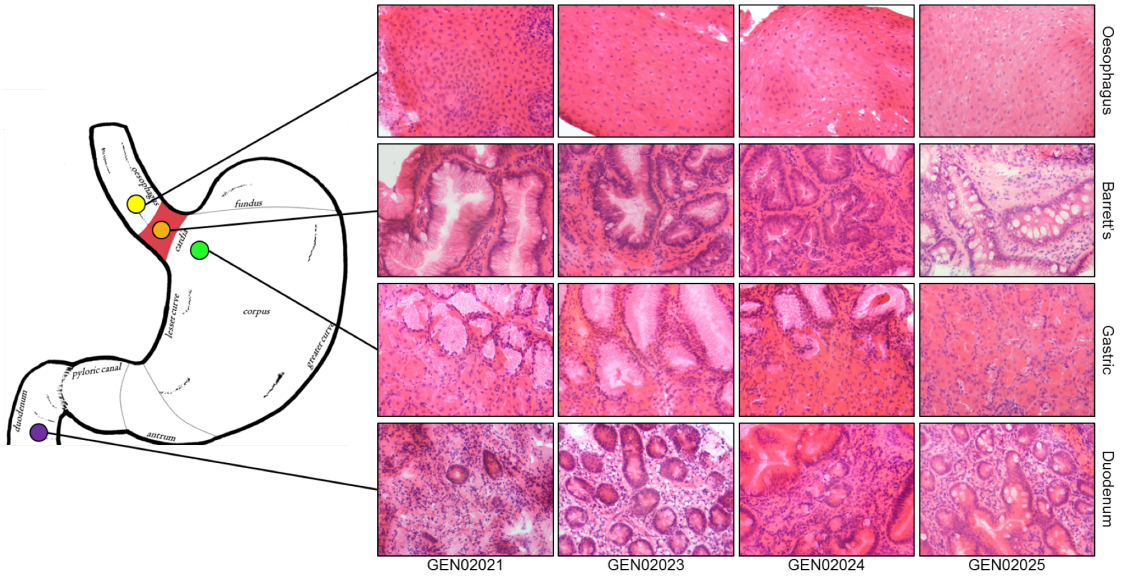


Figure A.1.4: An anatomical illustration (left) of the human upper gastrointestinal tract, where the sampling points for the endoscopic biopsies are indicated by color (Barrett's: orange, oesophagus: yellow, gastric: green, duodenum: purple). The Barrett's lesions typical anatomical location is indicated in red. Lines from sampling points for endoscopic biopsies point to the histology performed by Drs. Michael White and Richard Owen. A panel is present for each patient and tissue that presented with Barrett's oesophagus.

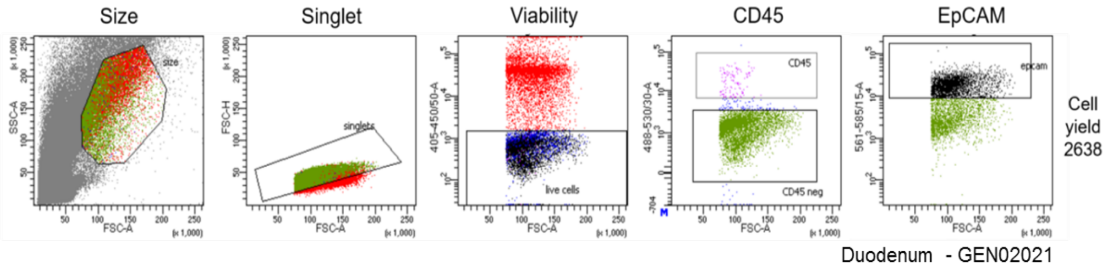


Figure A.1.5: A exemplar output of the FACS sorting protocol performed by Drs. Richard Owen and Michael White in order to isolate epithelial single cells from primary biopsies. The output shown is the result obtained from the duodenal tissue biopsied from patient GSE02021. Cells were gated fro size, viability, CD45-, and Epcam+ traits. Cells were sorted into 96-well plates.

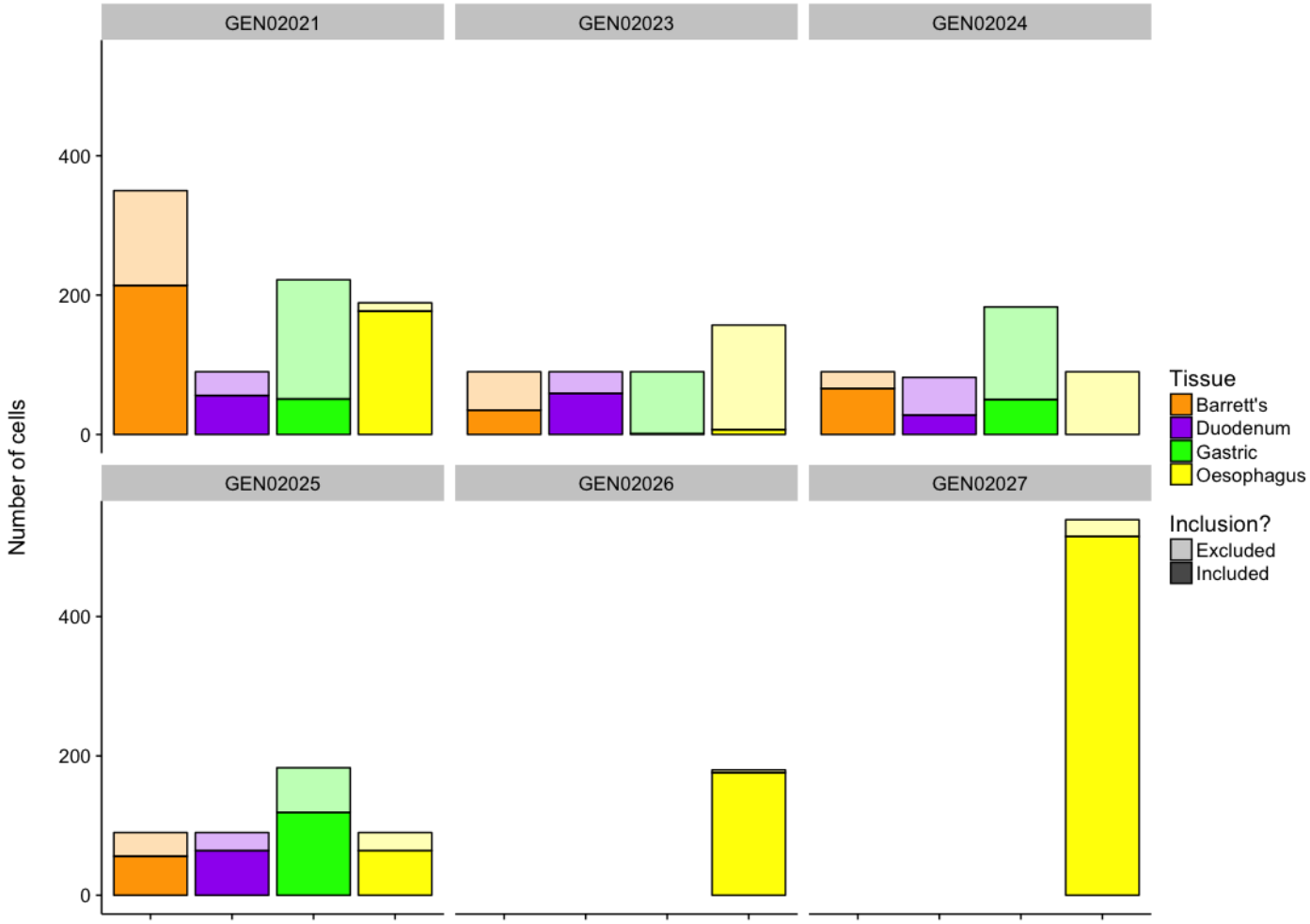


Figure A.1.6: Barplots illustrate the final inclusion and exclusion of cells in downstream analysis by tissue (annotated by color) and patient (panels). Each bar indicates the number of cells either excluded (translucent) or included (opaque) due to quality control filtering. The tissue types are annotated by color.

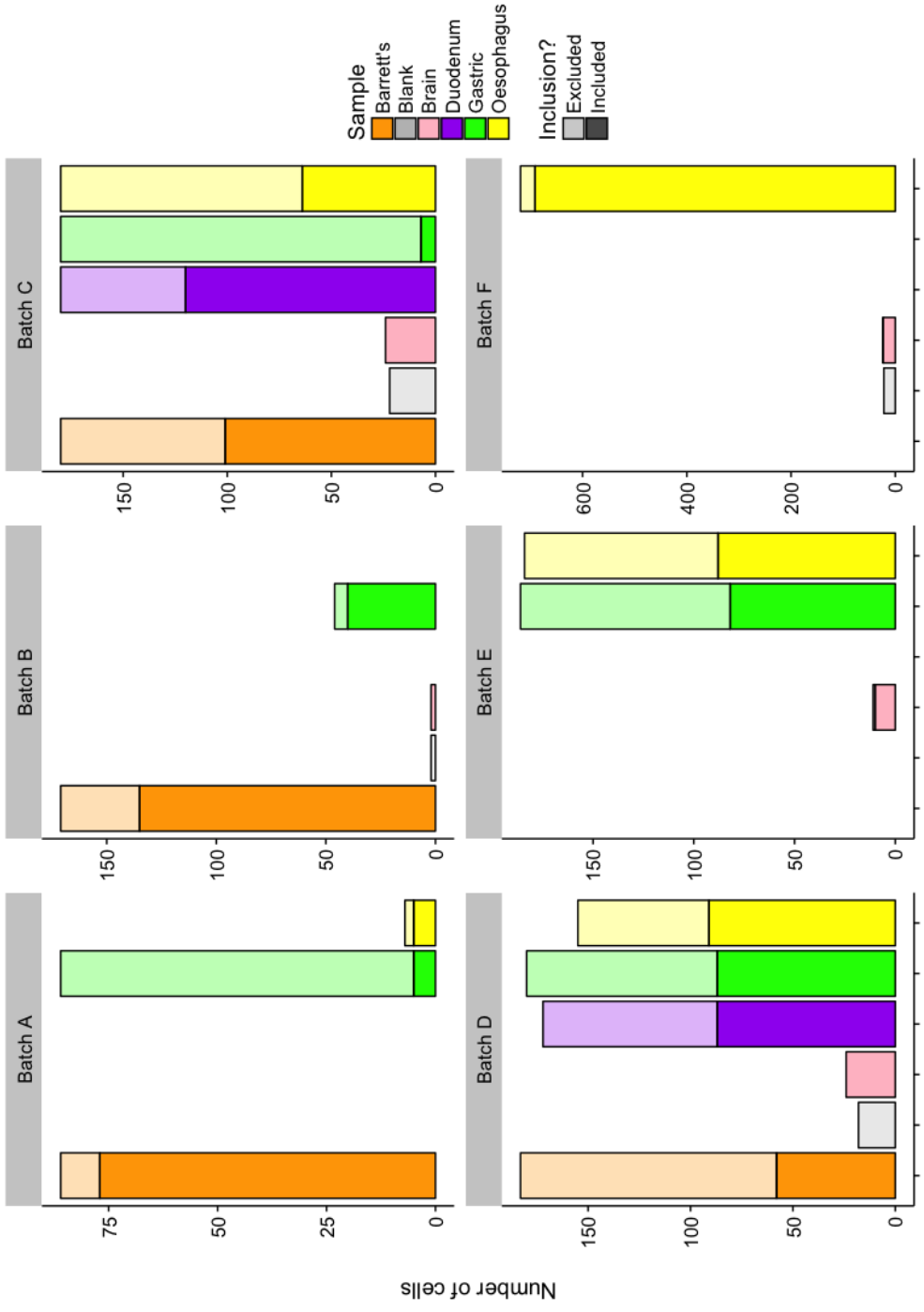


Figure A.1.7: Barplots illustrate the final inclusion and exclusion of cells in downstream analysis by tissue (x-axis) and batch (panels). Each bar indicates the number of cells either excluded (translucent) or included (opaque) due to quality control filtering. The tissue types are annotated by color.

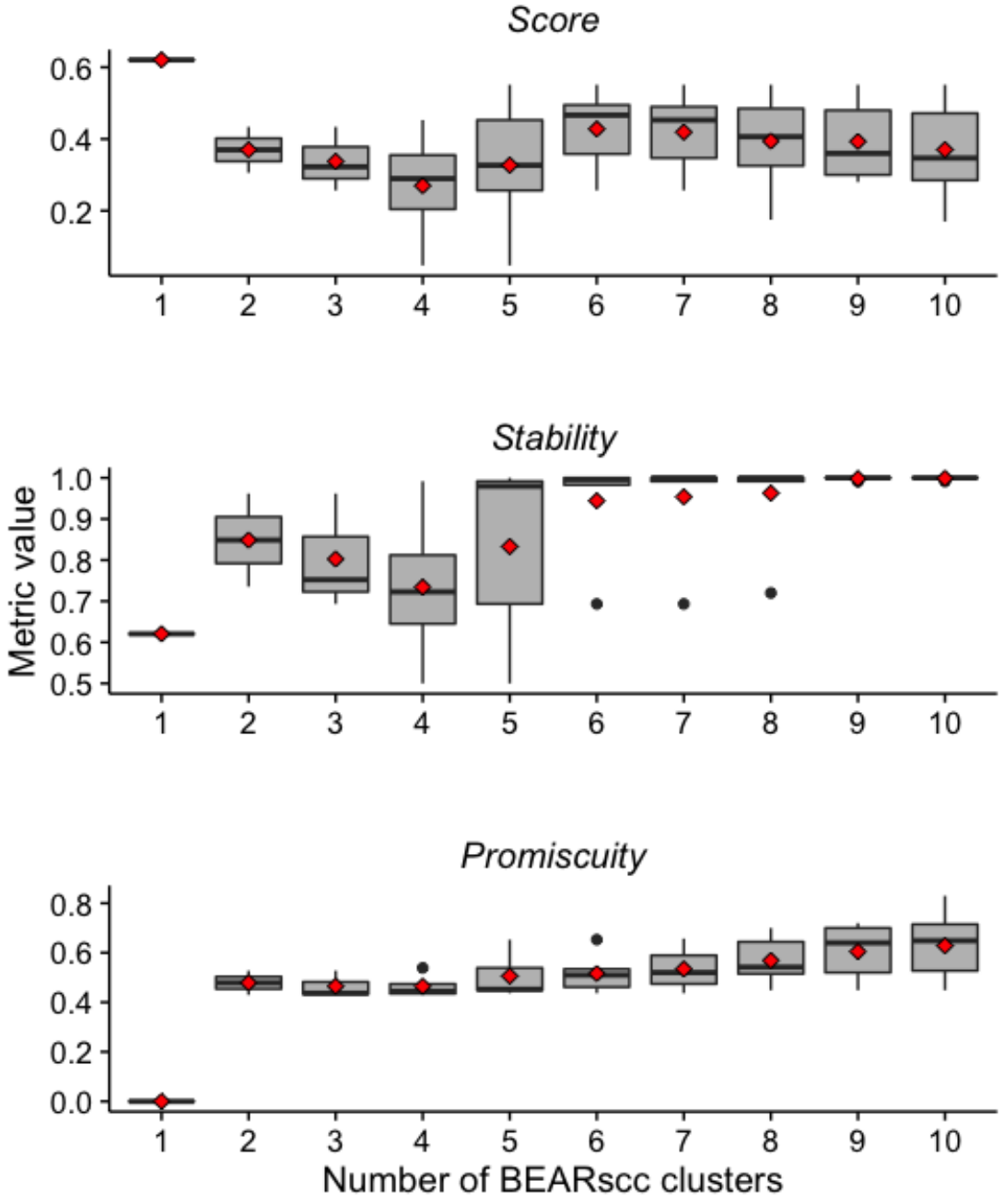


Figure A.1.8: Boxplots display the distribution of BEARscc cluster score for each hierarchical clustering on the noise consensus of QC-filtered samples in cluster 6 at some cluster number (x-axis). Each panel displays one of the three BEARscc metrics: score, stability, or promiscuity. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The red diamonds indicate the mean BEARscc clustering score across clusters for each hierarchical clustering of the noise consensus at various cluster numbers.

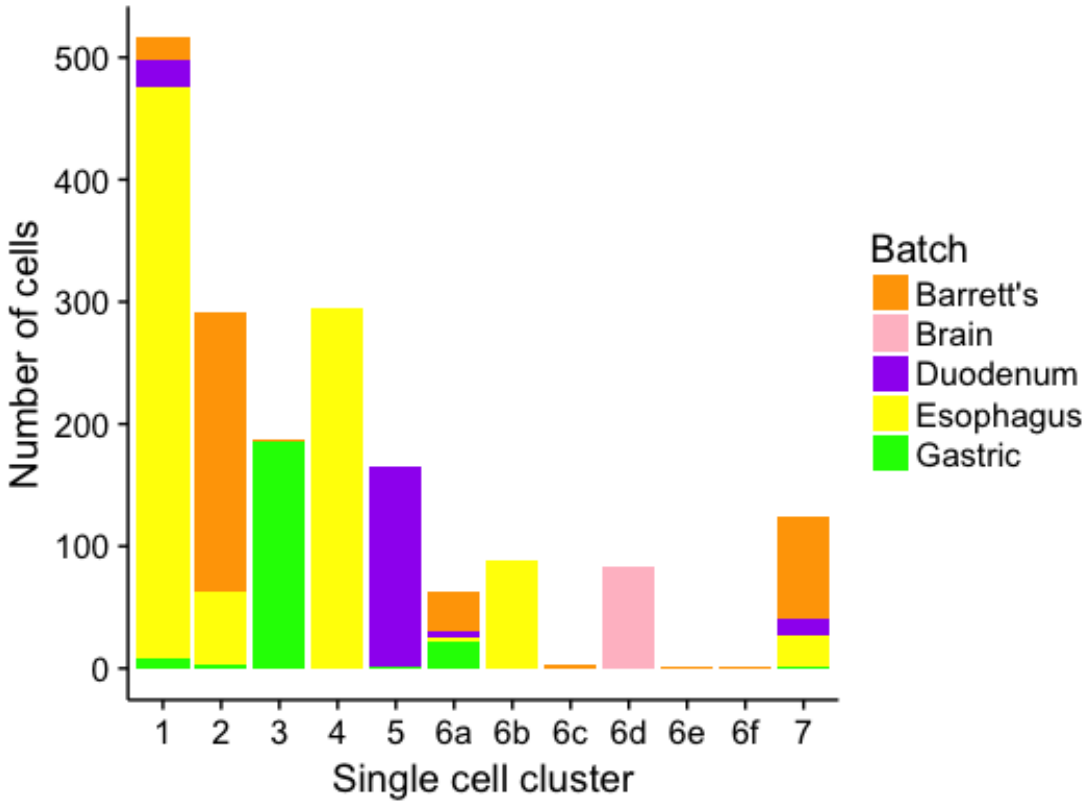


Figure A.1.9: Some single cell clusters from all QC-filtered samples are dominated by a single tissue type and others are an admixture of tissues. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by the tissue or sample from which the single cell samples were isolated.

A.1.3 Esophageal cluster interpretation

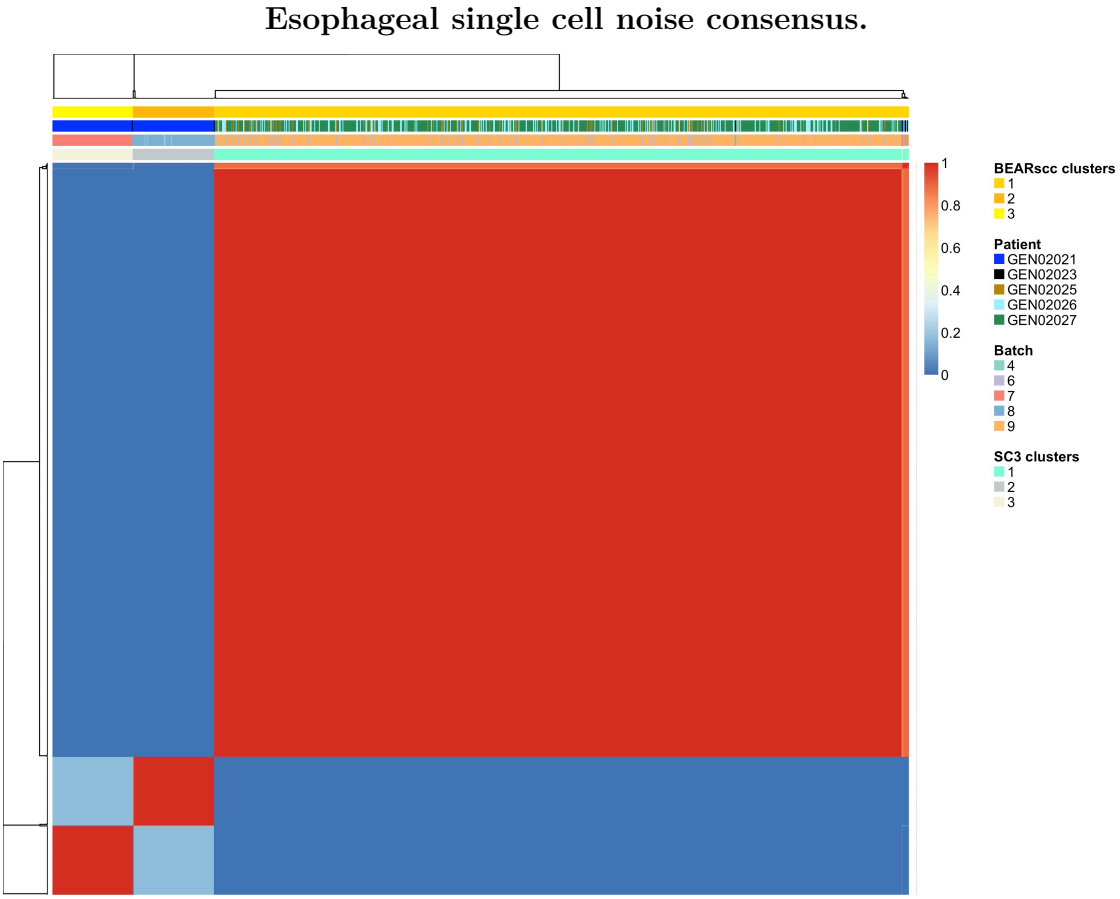


Figure A.1.10: The noise consensus of the proximal normal esophageal cells. Heatmap displays the noise consensus from SC3 clustering with $k = 3$ of 100 simulated technical replicates of cells isolated from endoscopic biopsies of the proximal squamous esophagus. Elements of the heatmap indicate the association frequencies of cells ordered by complete hierarchical clustering of the noise consensus. The 3 BEARscc clusters previously selected using the score metric distribution of clusters are annotated by color above the heatmap. Patient, batch and the initial SC3 clustering with $k = 3$ clusters of the observed data are also annotated by color above the heatmap.

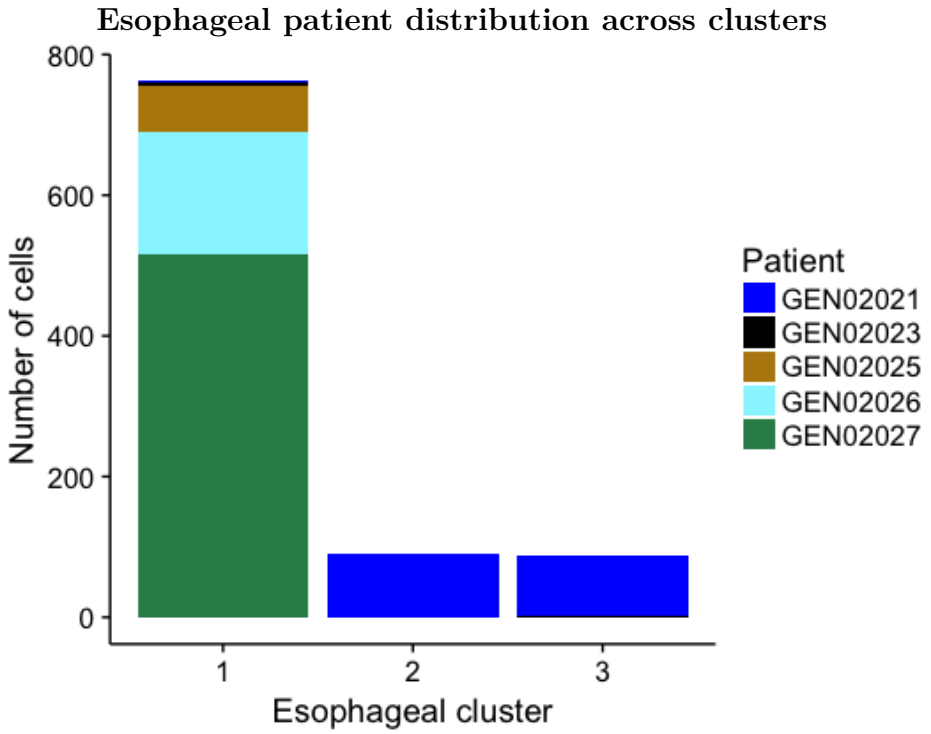


Figure A.1.11: The esophageal clusters are heavily confounded by patient source. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

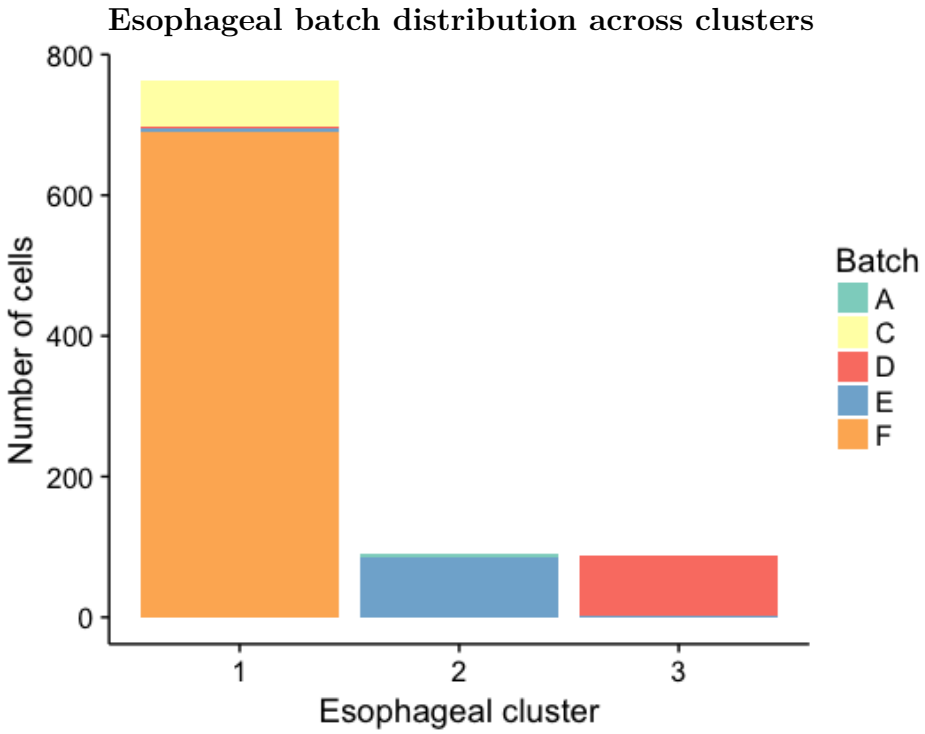


Figure A.1.12: The esophageal clusters are heavily confounded by batch. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

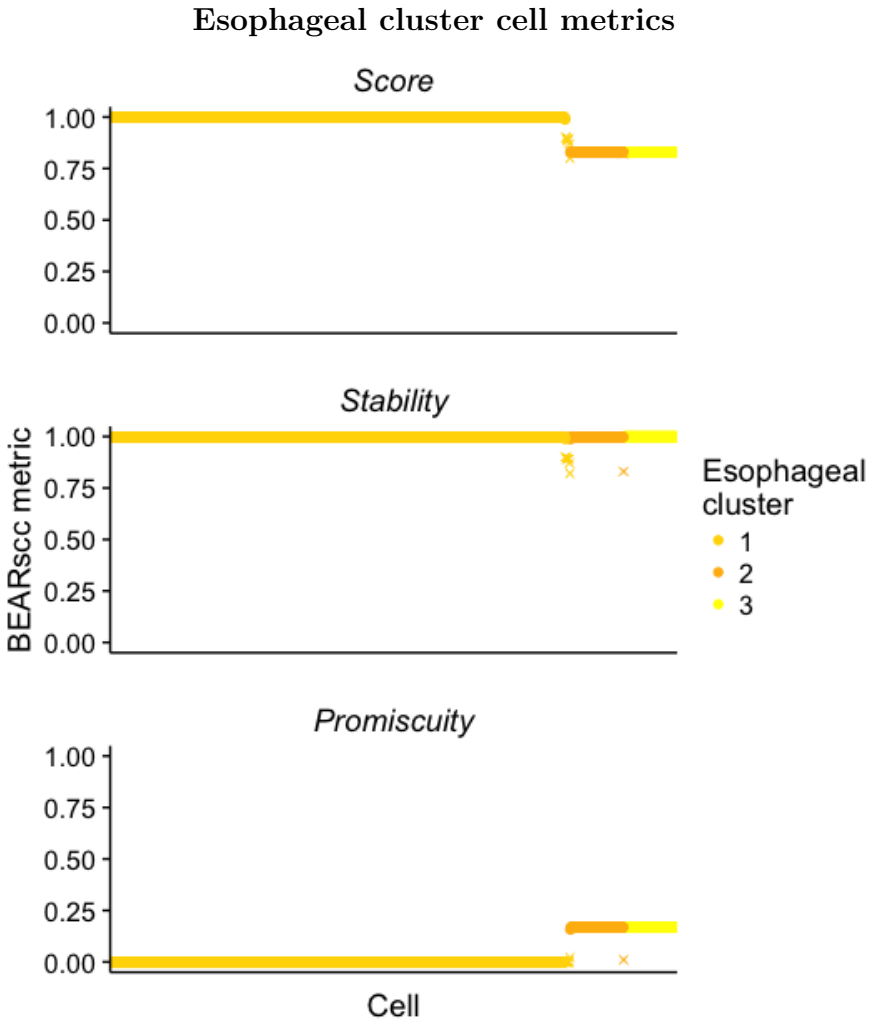


Figure A.1.13: Scatterplots display the BEARscc metrics (panels: score, stability, promiscuity) for each esophageal cell (x-axis) ordered by the esophageal cluster to which it belongs and the BEARscc score metric. Esophageal single cell clusters are annotated by color. Inclusion for differential expression analysis is indicated by dots, while exclusion is indicated by crosses.

Brief explanation of interpretation Among the differentially expressed genes, I found that several genes known to be important in esophageal biology were present for each cluster (Figure A.1.14). Specifically, clusters 1 and 2 exhibited clear squamous character with high gene expression of keratins, KRT5 and KRT15. Furthermore, these keratin 5 is associated with basal layers of the esophageal squamous epithelium, and the basal character of these cells is further supported by the consistent expression of TP63 across cells in clusters 1 and 2. Interestingly, a number of cluster 1 cells express KRT4/KRT13, which denotes more differentiated squamous epithelia in the esophagus. In contrast, cluster 2 cells generally do not express KRT4 and have relatively low KRT13 expression, and cells from cluster 2 are unique in their expression of inflammatory genes, such as IF144L and VCAM1. Therefore, I concluded that while both cluster 1 and 2 represented basal squamous epithelium, cluster 1 cells were either less stressed or more differentiated than cluster 2 cells. The presence of inflammatory markers in cluster 2 led us to conclude that the former hypothesis was more likely.

Cluster 3 was extremely unique its profile of differentially expressed genes; a keratin profile was completely absent in these cells. Furthermore, these cells expressed TFF3 and MUC6, which have been reported to be expressed only in cells found in the rare esophageal submucosal glandular ducts. Drs. Richard Owen and Michael White verified the expression differences between esophageal squamous cells and esophageal submucosal glandular ducts with immunohistochemistry (Supplementary Figure A.1.16). It should be noted that cluster 3 contained cells from only two patients, GEN02021 and GEN02023, with the vast majority of cells originating form the former. However, given that esophageal submucosal glandular ducts are rare (especially so in patients without Barrett's esophagus), and the logistical problem of under-sampling of the proximal normal squamous esophageal epithelium in severeal patients, it is conceivable that the patient bias is the result of sampling. Moreover, explaining these observations with haplotype differences between patients seems implausible given the radical difference expression profiles and the presence of the cluster 3 profile in more than one patient. I proceeded

with the understanding that clusters 1 and 2 consisted of unstressed and stressed basal squamous esophageal epithelium, respectively, and cluster 1 consisted of cells from esophageal submucosal glandular ducts.

Curated list of differentially expressed genes in esophageal clusters

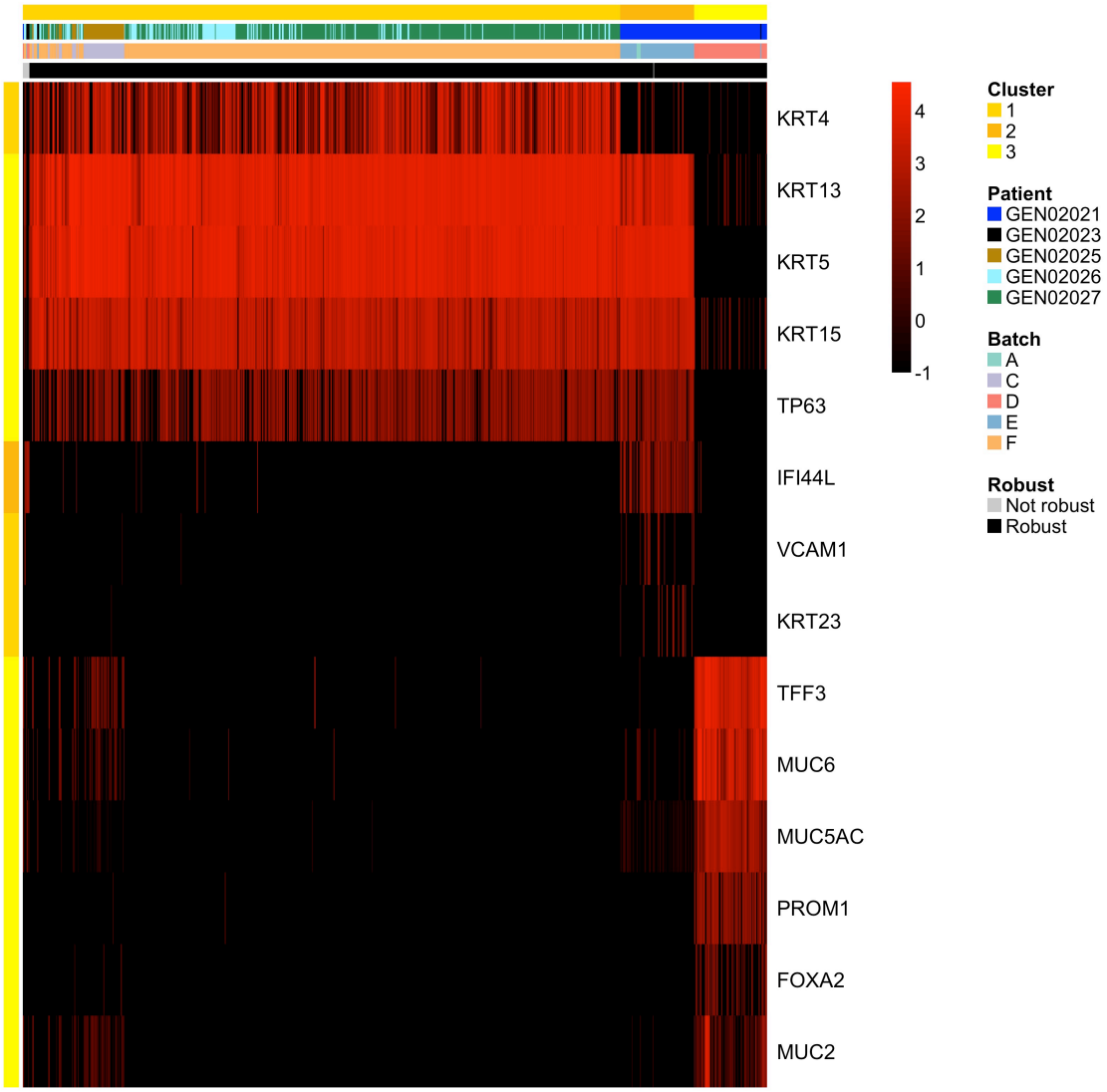


Figure A.1.14: Heatmap displays gene expression in esophageal cells in log10(TPM+0.1) for select differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

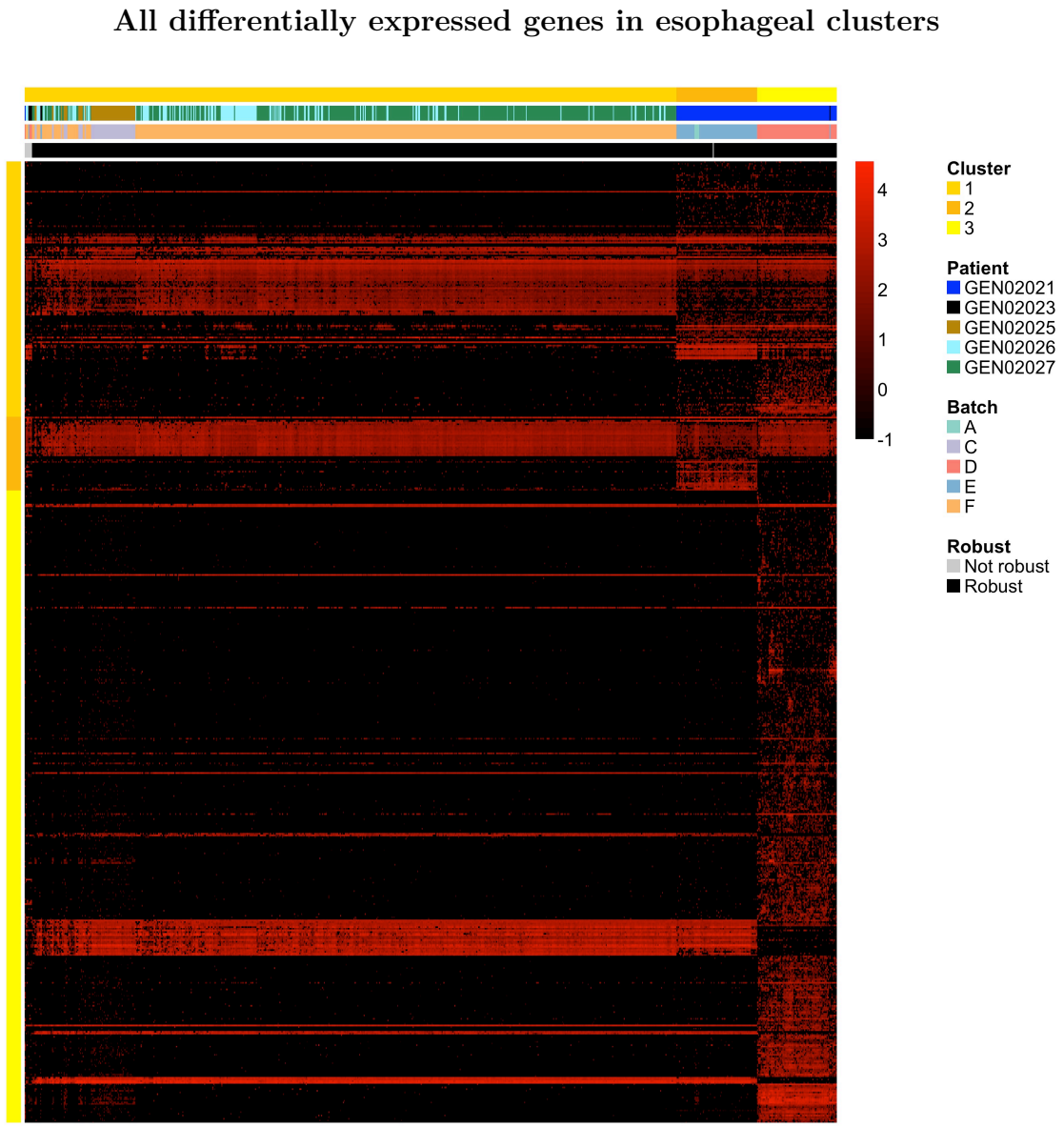


Figure A.1.15: Heatmap displays gene expression in esophageal cells in log10(tpm+0.1) for differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

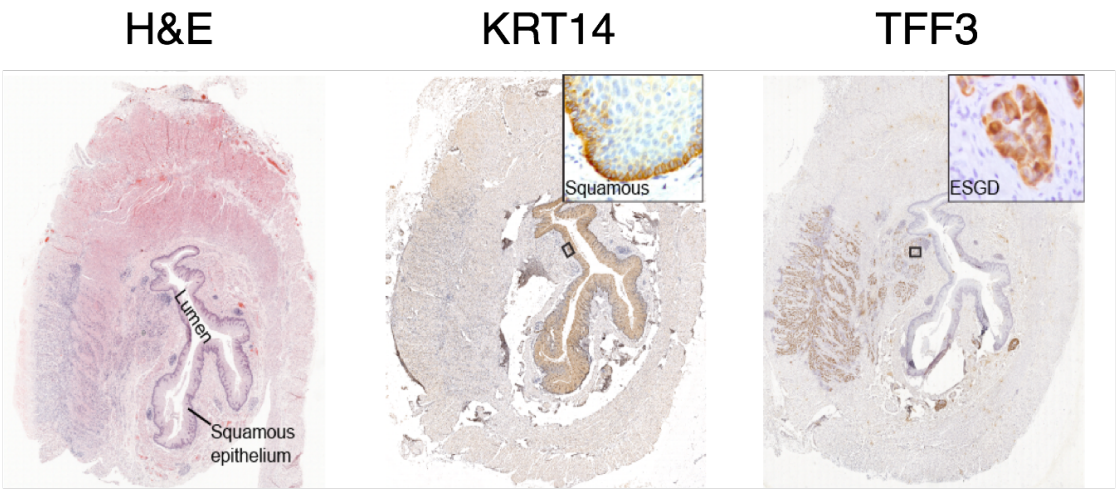


Figure A.1.16: Representative image of hematoxylin and eosin staining of normal esophagus taken from an esophagectomy specimen (left). immunohistochemical staining of KRT14 (center) and TFF3 (right) in adjacent sections from the same specimen with enlargements of representative positively stained regions. Immunohistochemistry and legend are the work of Dr. Michael White.

A.1.4 Gastric cluster interpretation

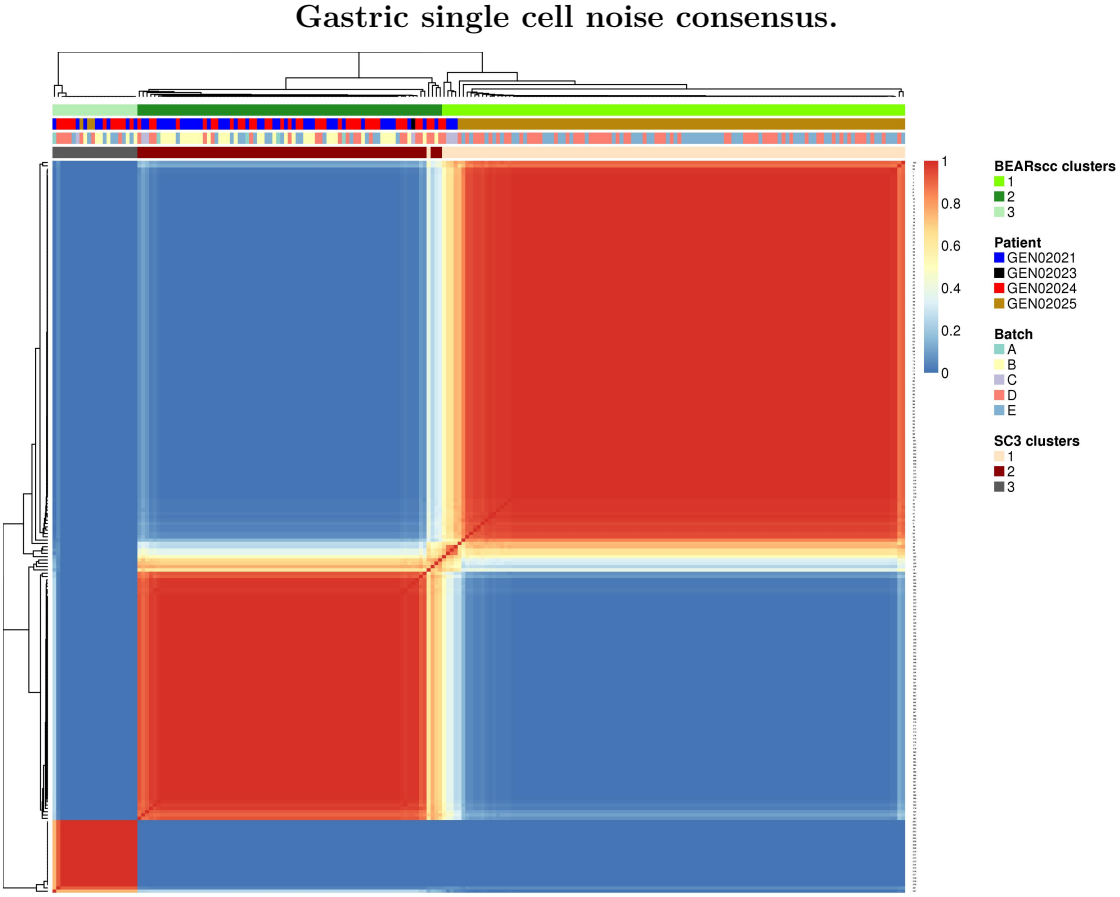


Figure A.1.17: The noise consensus of the gastric cells. Heatmap displays the noise consensus from SC3 clustering with $k = 3$ of 100 simulated technical replicates of cells isolated from endoscopic biopsies of the proximal stomach. Elements of the heatmap indicate the association frequencies of cells ordered by complete hierarchical clustering of the noise consensus. The 3 BEARscc clusters previously selected using the score metric distribution of clusters are annotated by color above the heatmap. Patient, batch and the initial SC3 clustering with $k = 3$ clusters of the observed data are also annotated by color above the heatmap.

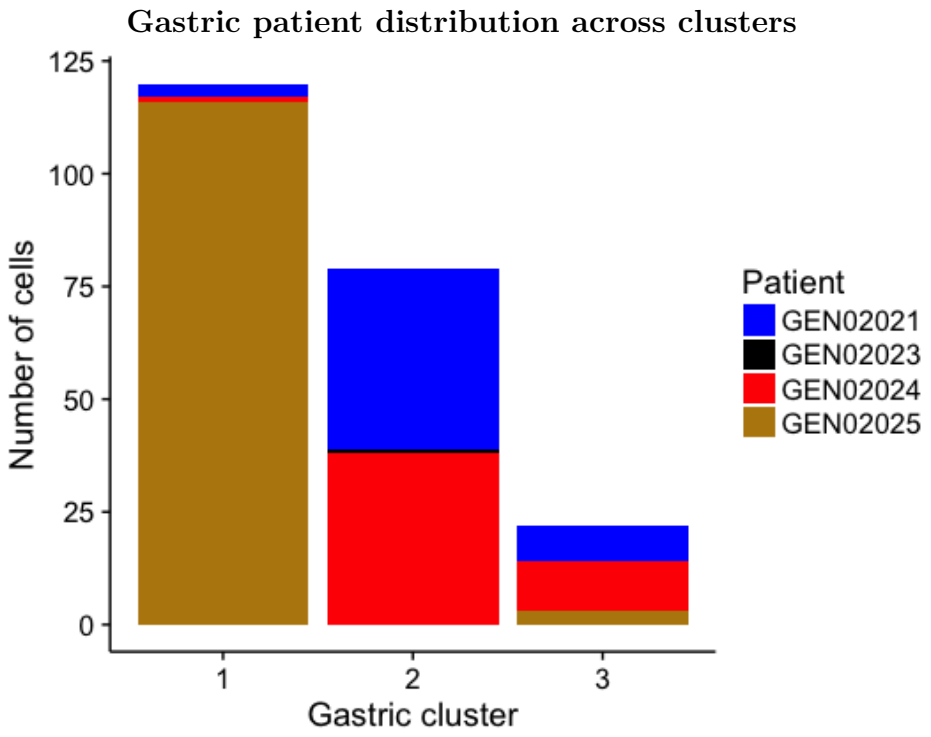


Figure A.1.18: The gastric clusters appear to be somewhat confounded by patient source. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

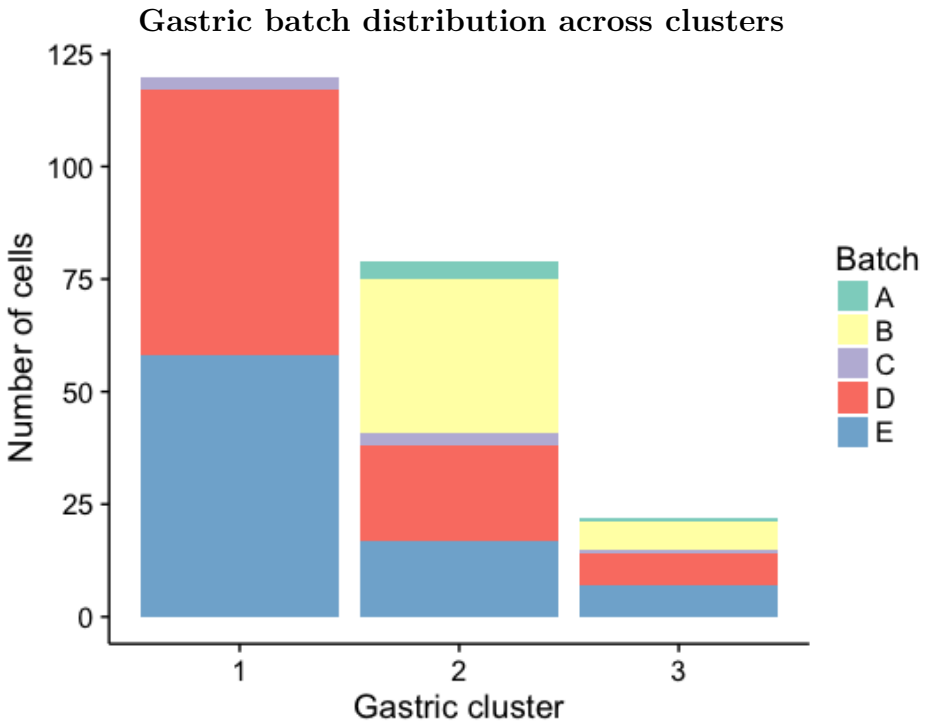


Figure A.1.19: The gastric clusters do not appear to be confounded by batch. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by batch from which the single cell samples were isolated.

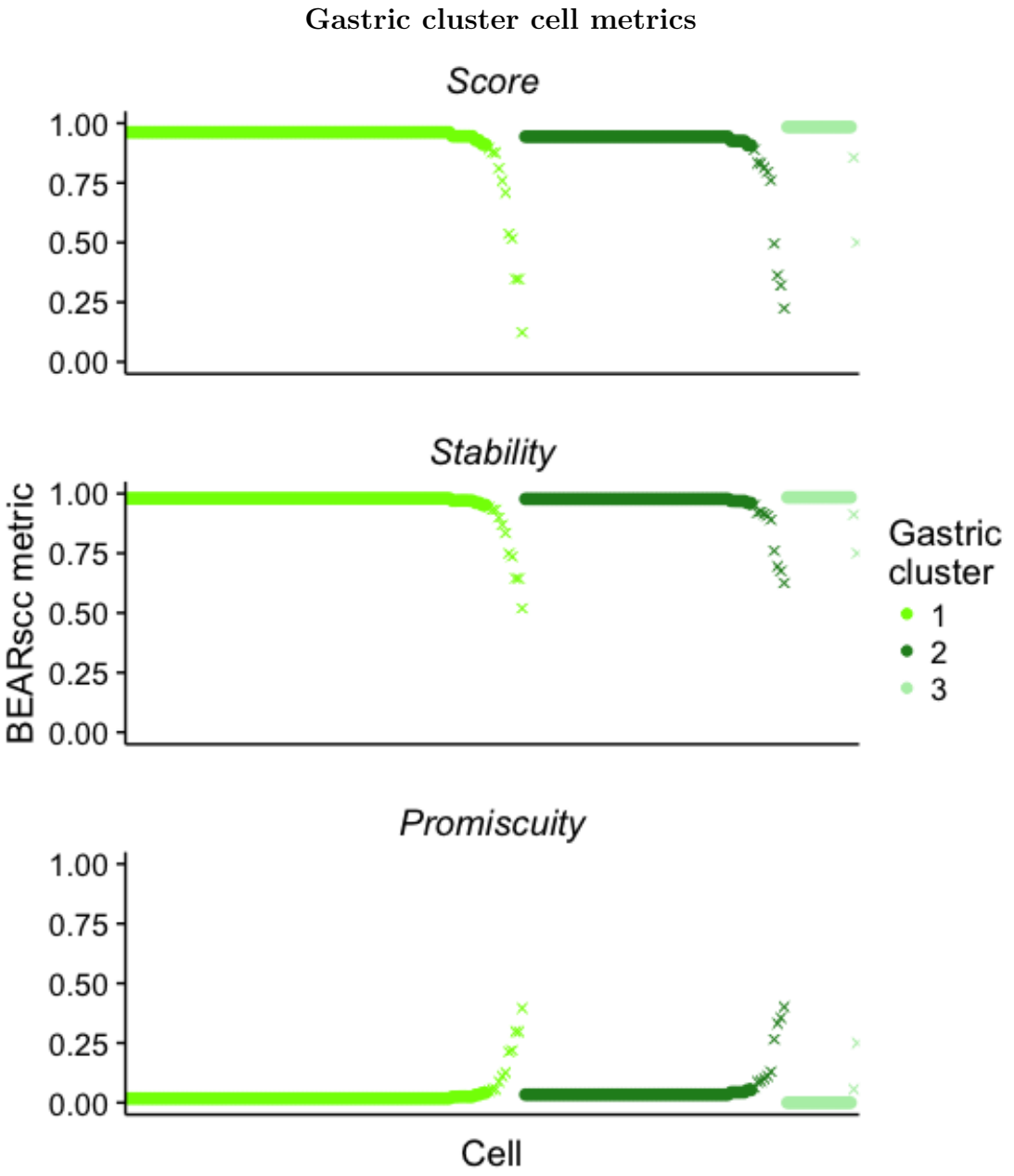


Figure A.1.20: Scatterplots display the BEARscc metrics (panels: score, stability, promiscuity) for each gastric cell (x-axis) ordered by the gastric cluster to which it belongs and the BEARscc score metric. Gastric single cell clusters are annotated by color. Inclusion for differential expression analysis is indicated by dots, while exclusion is indicated by crosses.

Brief explanation of interpretation The principal distinction I observed in the gastric BEARscc clusters existed between cluster 3 and clusters 1 and 2 (Figure A.1.21). Cluster 3 cells were chromogranin-A (CHGA) positive, and generally also expressed secretogranin (SCGN) and chromogranin-B (CHGB), while generally exhibiting low expression for mucin genes, such as MUC5AC and MUC4. Chromogranin-A is a canonical marker for entero-endocrine cells, and the presence of secretogranin and chromograinin-B futher support the conclusion that cluster 3 cells are, in fact, gastric entero-endocrinocytes. In contrast, cells in clusters 1 and 2 exhibited little to no chromgranin-A, secretogranin, and chromogranin-B.

The distinction between cluster 1 and 2 was much less clear. Indeed, I previously observed in the patient distribution that cluster 1 and 2 may be subject to patient genotype biases, and upon examining the differential expression results (Supplementary Table A.10), I observe that human leukocyte antigen (HLA) molecules are among the most significant results with the highest magnitude of fold change in expression. The HLA postive hits in differential expression analysis coupled with the subtle difference in marker genes between clusters 1 and 2 suggests that the only biological distinction between these two clusters may be genotype.

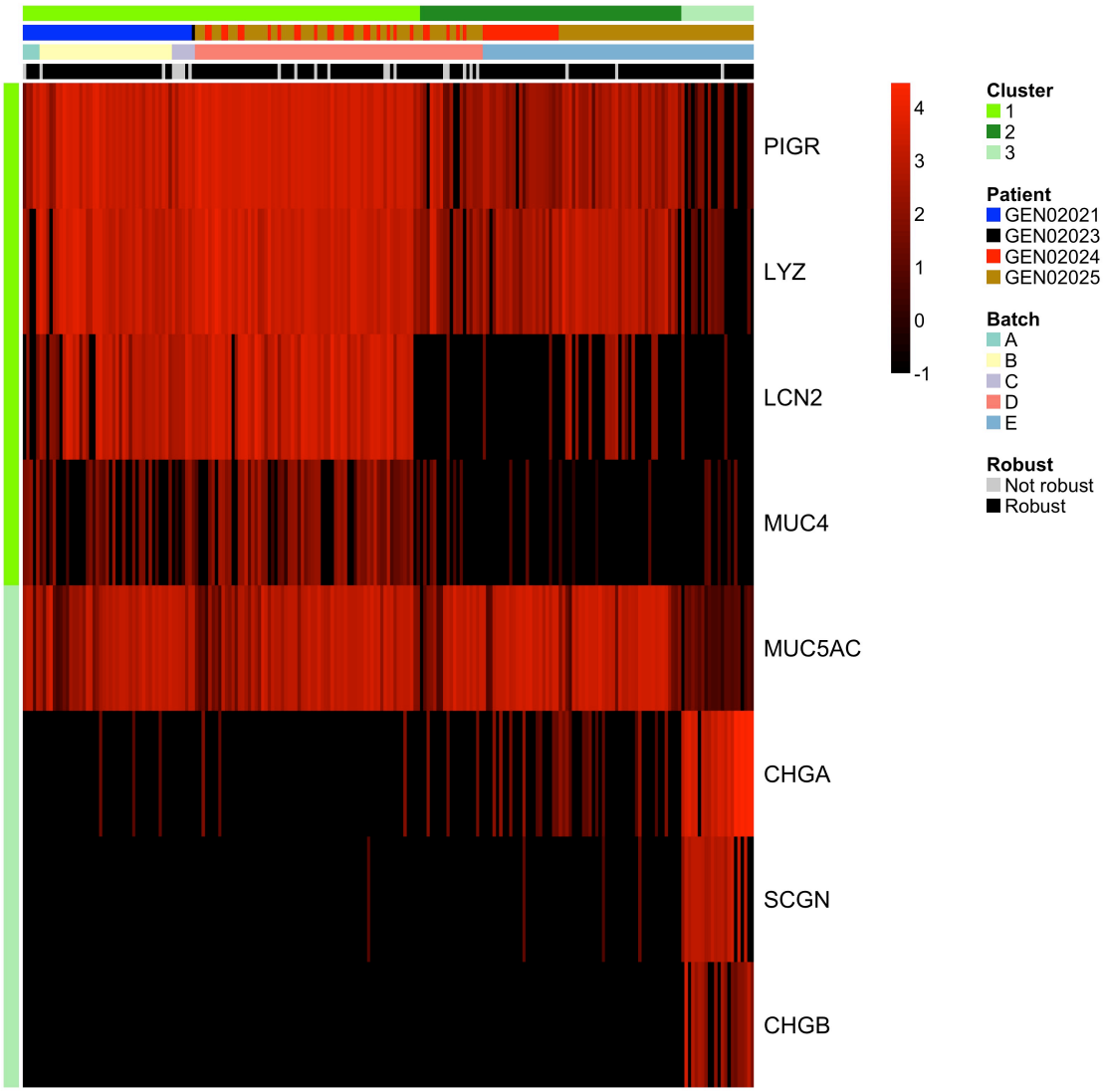


Figure A.1.21: Heatmap displays gene expression in gastric cells in log10(TPM+0.1) for select differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

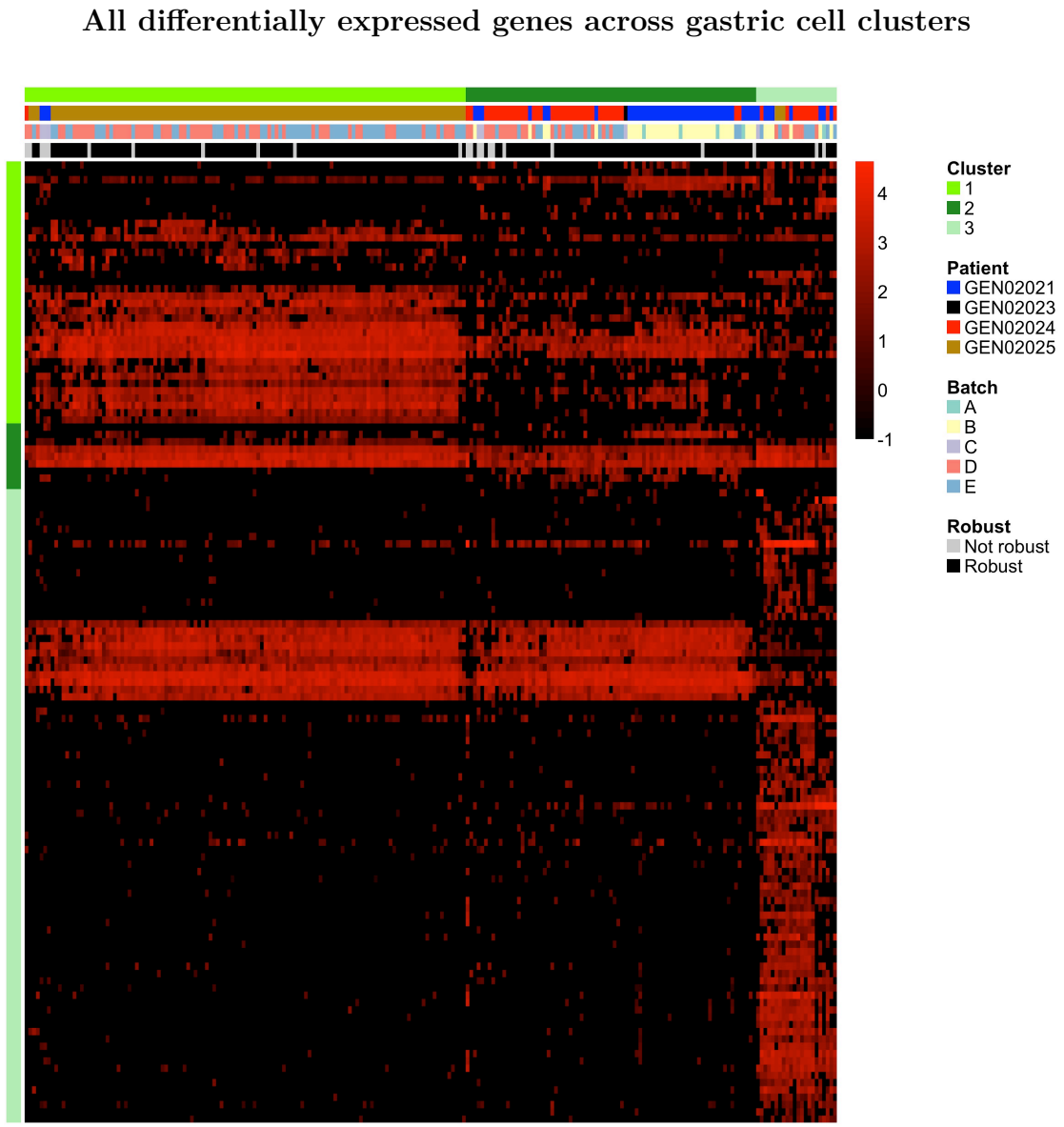


Figure A.1.22: Heatmap displays gene expression in gastric cells in log10(tpm+0.1) for differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

Closer examination of gastric cluster expression I examined several marker genes more closely in order to refine the interpretation of gastric clusters (Figure A.1.23), I observed that subtle differences in the cells populations do exist. The slightly higher expression of polymeric immunoglobulin receptor (PIGR), lysozyme (LYZ), mucin-4 (MUC4), and lipocalin 2 (LCN2) in cluster 1, and the subtle increase in mucin-5AC (MUC5AC) in cluster 2 could point to a genuine difference in cell type biology independent of patient genotype. Specifically, the properties of cluster 1 are more characteristic of mucous neck cells, while cluster 2 is more characteristic of surface mucous cells (Hoffmann 2015). Perhaps, the clear distinction between these groups was masked by genotype or perhaps the differentiation of these two gastric cell types is downstream of mRNA regulation. In either case, our experiment did not have the power to definitively differentiate the two cell types. Therefore, I concluded that both cluster 1 and cluster 2 consisted of gastric mucous cells existing on a gradient between surface mucous cells and mucous neck cells, while cluster 3 consisted of entero-endocrinocytes.

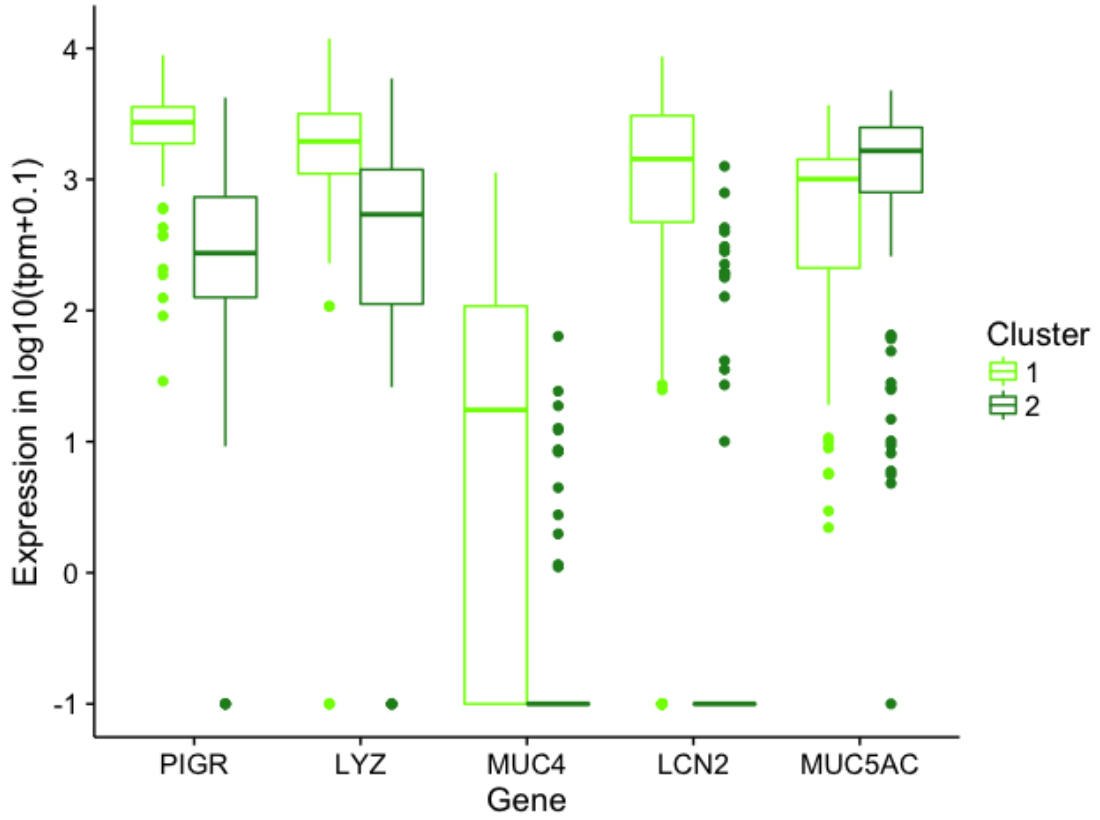


Figure A.1.23: Boxplots display the distribution of gene expression in log10(TPM+1) for select differentially expressed genes (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The color of the boxplots denote the gastric cluster to which the distribution belongs.

A.1.5 Duodenal cluster interpretation

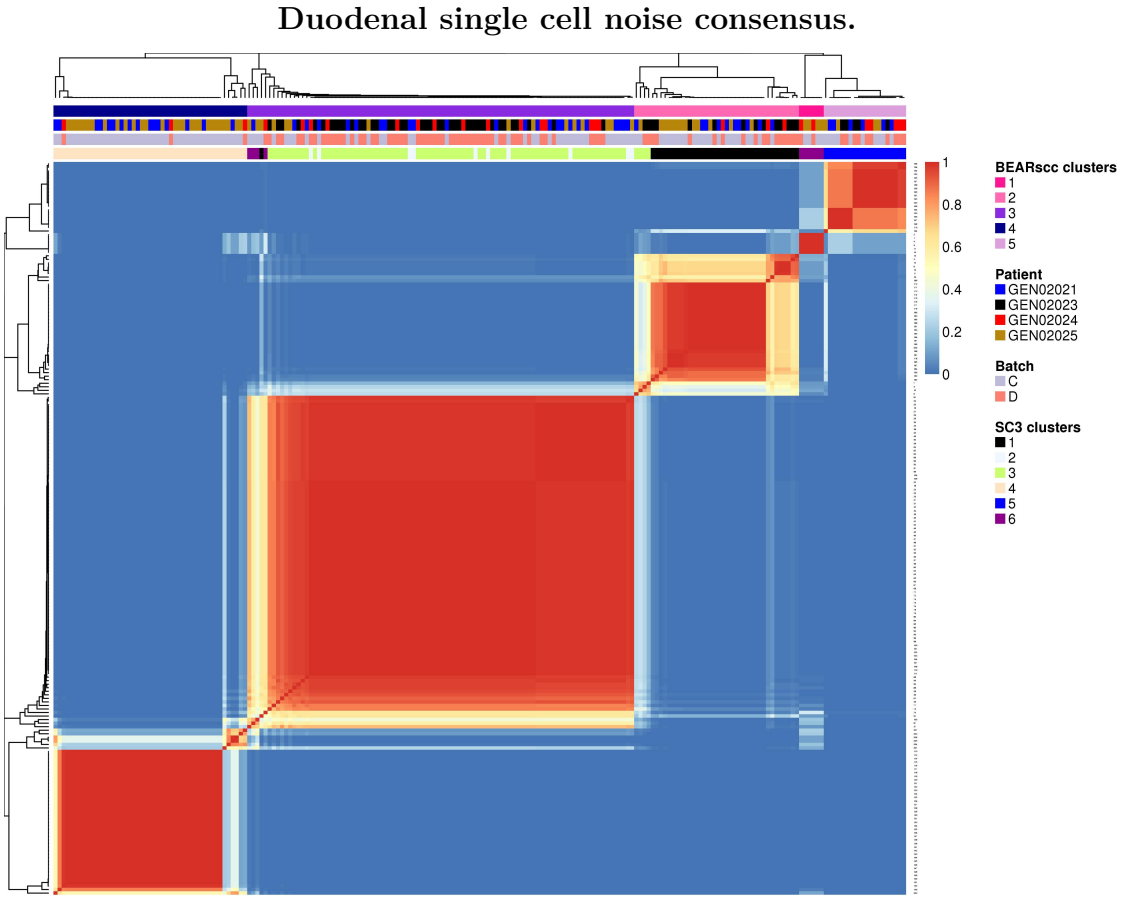


Figure A.1.24: The noise consensus of the duodenum. Heatmap displays the noise consensus from SC3 clustering with $k = 6$ of 100 simulated technical replicates of cells isolated from endoscopic biopsies of the duodenum. Elements of the heatmap indicate the association frequencies of cells ordered by complete hierarchical clustering of the noise consensus. The 5 BEARscc clusters previously selected using the score metric distribution are annotated by color above the heatmap. Patient, batch and the initial SC3 clustering with $k = 6$ clusters of the observed data are also annotated by color above the heatmap.

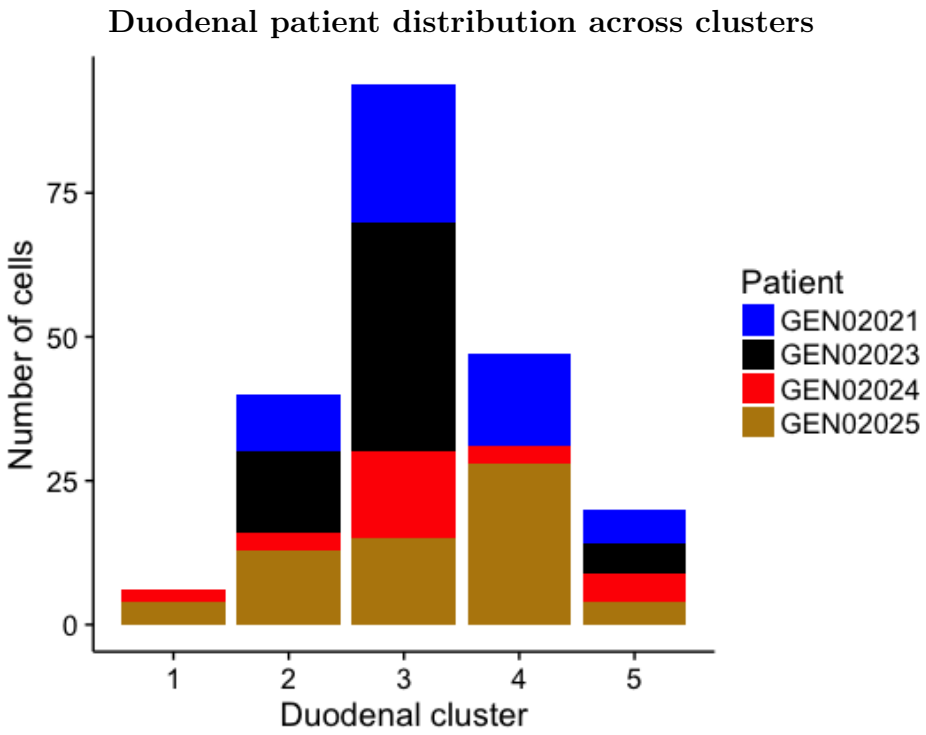


Figure A.1.25: The duodenal clusters do not appear to be confounded by patient source. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

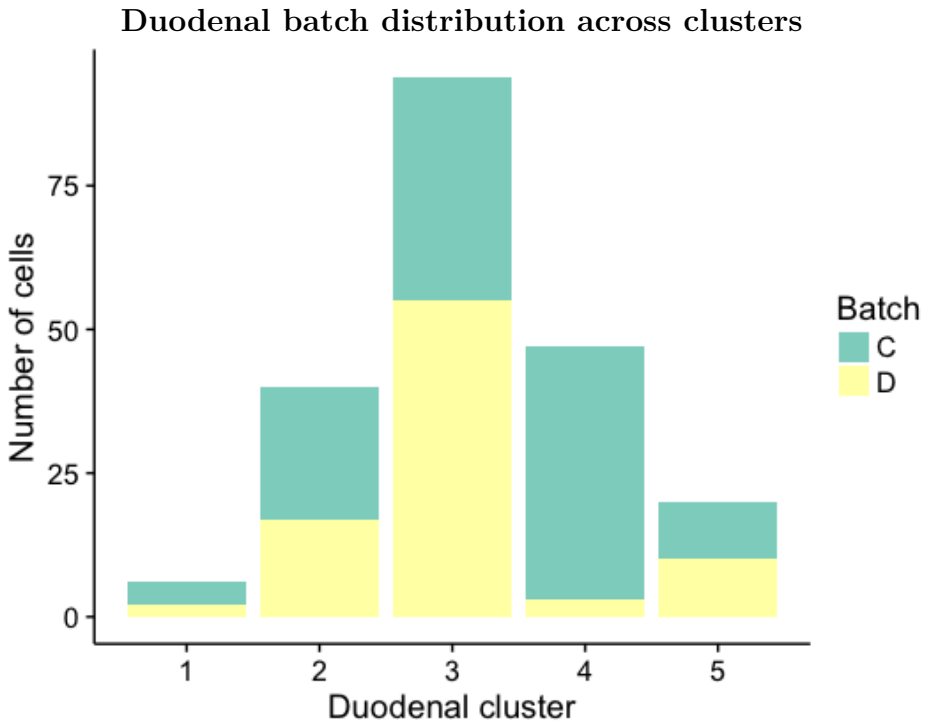


Figure A.1.26: The duodenal clusters do not appear to be confounded by batch. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by batch from which the single cell samples were isolated.

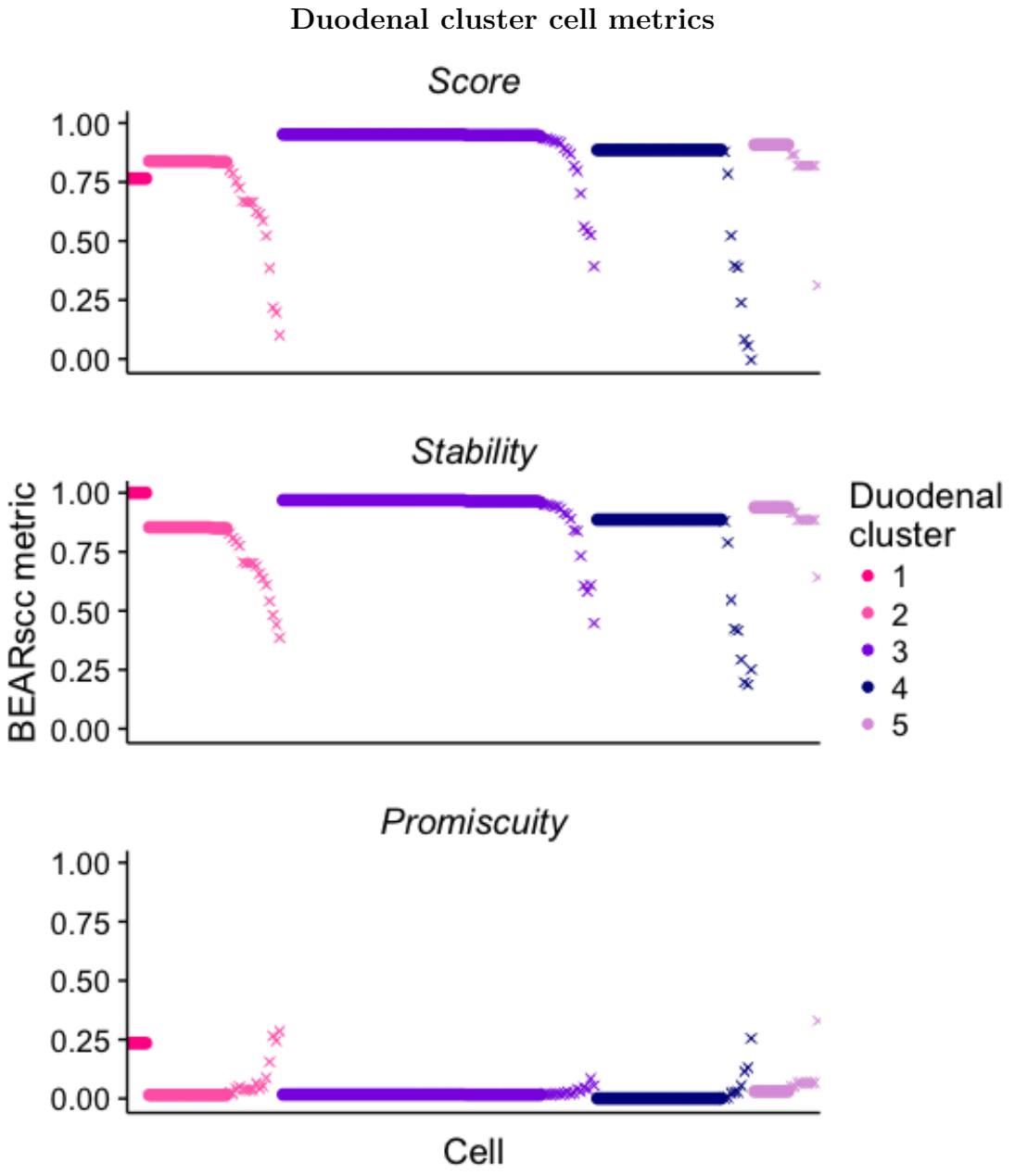


Figure A.1.27: Scatterplots display the BEARscc metrics (panels: score, stability, promiscuity) for each duodenal cell (x-axis) ordered by the Duodenal cluster to which it belongs and the BEARscc score metric. Duodenal single cell clusters are annotated by color. Inclusion for differential expression analysis is indicated by dots, while exclusion is indicated by crosses.

Brief explanation of interpretation Among the differentially expressed genes, I found that several genes known to be important in duodenal biology were present for each cluster (Figure A.1.28). Specifically, I found that cluster 1 was unique in its expression vimentin (VIM), a co-stimulatory immune checkpoint receptor (CD27), and a protein component of antibodies IgM and IgA (JCHAIN), which collectively suggested that these cells were either M-cells from Peyer's patches or mesenchymal non-epithelial cells.

I observed highly similar expression profiles between clusters 2 and 3 (Supplementary Figure A.1.29). In cluster 2, I observed that these cells were primarily distinguished by proliferation genes such as cyclin-dependent kinase 1 (CDK1), helicase expression (HELLS), and an E2F family transcription factor (E2F1). Cluster 3 was unique in that all genes identified as significant in differential analysis had lower expression than the rest of the duodenal samples; that is, cluster 3 was characterized by a lack of differentiated marker genes. Cluster 3 was still positive for genes ubiquitously expressed in the duodenum, such as VIL1 (data not shown), suggesting that cluster 3 consists of relatively unspecialized, even undifferentiated, duodenal epithelium.

In contrast, clusters 4 and 5 have clear specialized identities. Cluster 4 consists of cells with high intestine alkaline phosphatase (ALPI) and lactase (LCT). which are well-described marker genes for intestinal enterocytes. Cluster 5 consists of cells with high expression in secretory marker genes such as mucin-2 (MUC2) and trefoil factor 3 (TFF3). This suggests cluster 5 consists of secretory cells including goblet cells. I also observed that two genes, intelectin-1 (ITLN1) and a serine peptidase inhibitor (SPINK4), were among the most significant genes in differential expression analysis and associated strongly with the goblet cell marker MUC2. I hypothesized that these genes could be new markers for goblet cells, or even demarcate cells targeted to a goblet cell fate, but not yet differentiated to the point of expressing the canonical mucin profile found in goblet cells; however, further experimentation is necessary before any such conclusion may be drawn. Interestingly, REG4 (regenerating family member 4) is highly expressed in the cells

that were excluded from differential expression analysis due to a clear drop in score. REG4 has been described as a marker for entero-endocrinocytes of the intestine (Grün *et al.* 2015). Coupled with the observations in the noise consensus that these cells did not consistently associate with the goblet-like cells in cluster 5 to the point of being excluded from DE analysis and the marked visual difference in expression profiles between these groups in the expression profile of differentially expressed genes, I hypothesized that these cells may be entero-endocrinocytes.

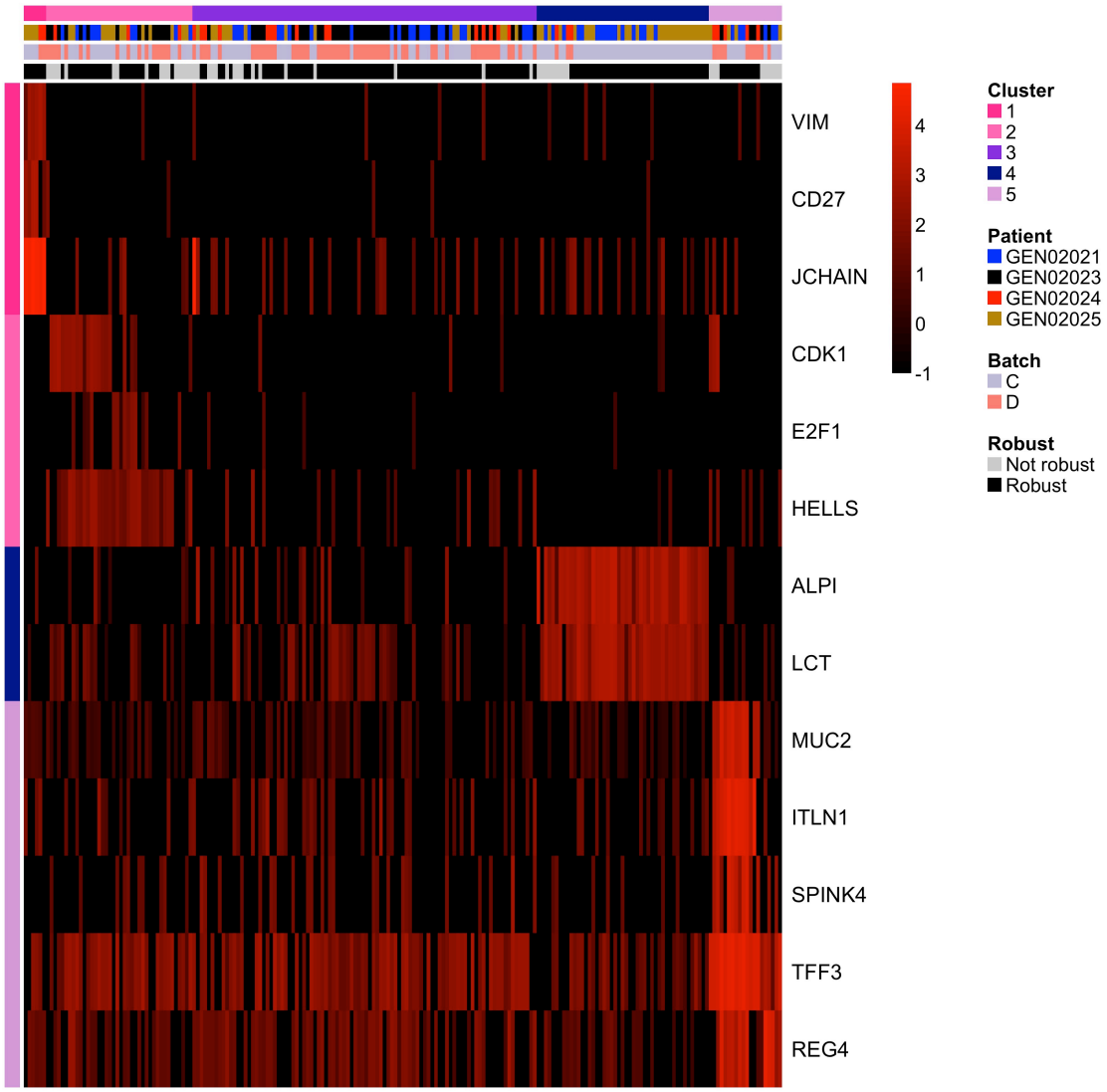


Figure A.1.28: Heatmap displays gene expression in duodenal cells in log10(TPM+0.1) for select differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

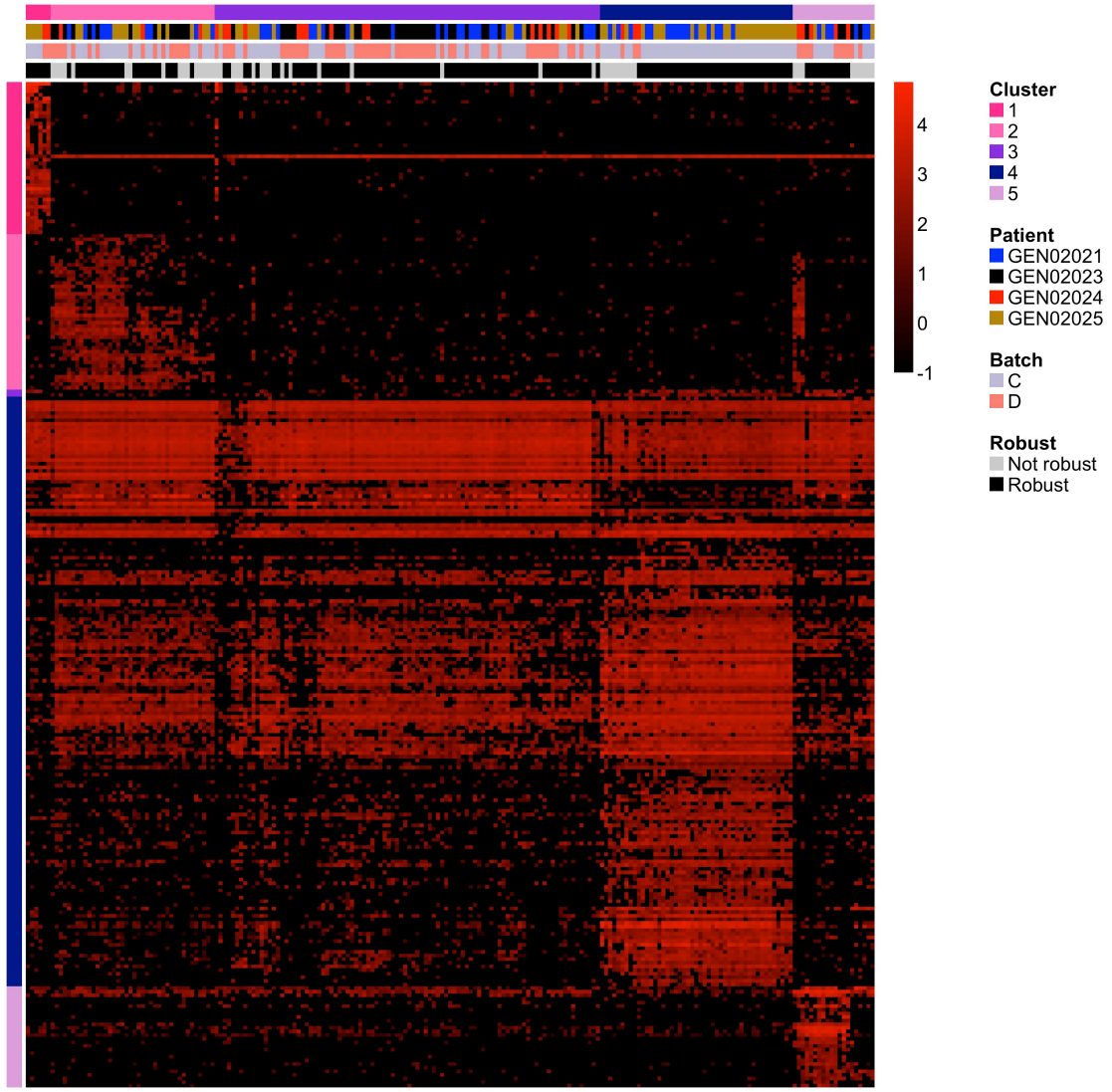


Figure A.1.29: Heatmap displays gene expression in duodenal cells in log10(tpm+0.1) for differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

Closer examination of duodenal cluster expression To explore further the hypothesis that cluster 3 consisted of less well-differentiated and stem cells, I examined the expression of the known intestinal stem cell marker, LGR5, in the 5 duodenal clusters (Figure A.1.30). I observed that clusters 2 and 3 contained 28 of 31 cells with any observed LGR5 expression, and cluster 3 contained the four

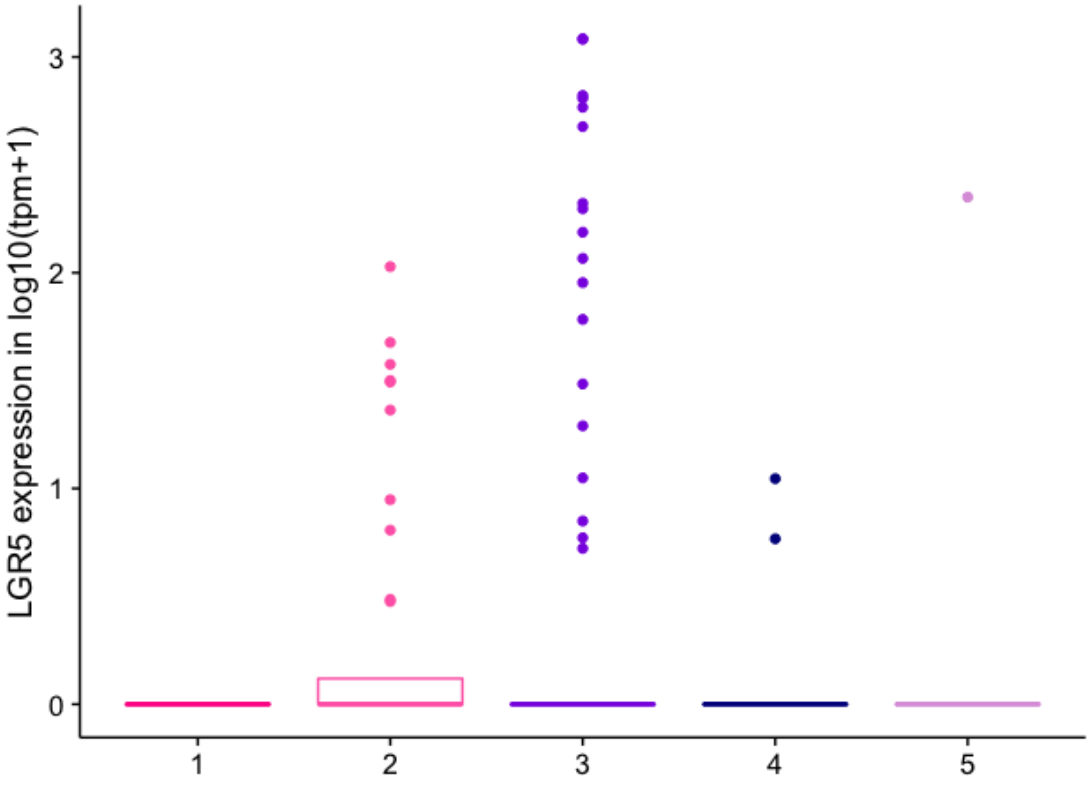


Figure A.1.30: Boxplots display the distribution of LGR5 expression in log10(TPM+1) by duodenal BEARscc cluster (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. BEARscc duodenal clusters are also coded by color.

cells with the highest LGR5 expression amongst the sequenced duodenal libraries. From these observations, I concluded that clusters 2 and 3 consist of generally less well-differentiated cells from the duodenal crypt and contain the duodenal stem cells.

To test the hypothesis that cluster 5 consisted of a subset of entero-endocrine cells, I examined the expression of chromogranins, CHGA and CHGB, and secretogranin, SCGN, and REG4 of cells in cluster 5 (Figure A.1.31). I observed that the cells from cluster 5 excluded from differential expression analysis, did indeed have high expression of CHGA, CHGB, and SCGN as compared to cells that were included. Therefore, I concluded that cluster 5 predominantly consisted of secretory intestinal cells of two types: entero-endocrine and goblet cells.

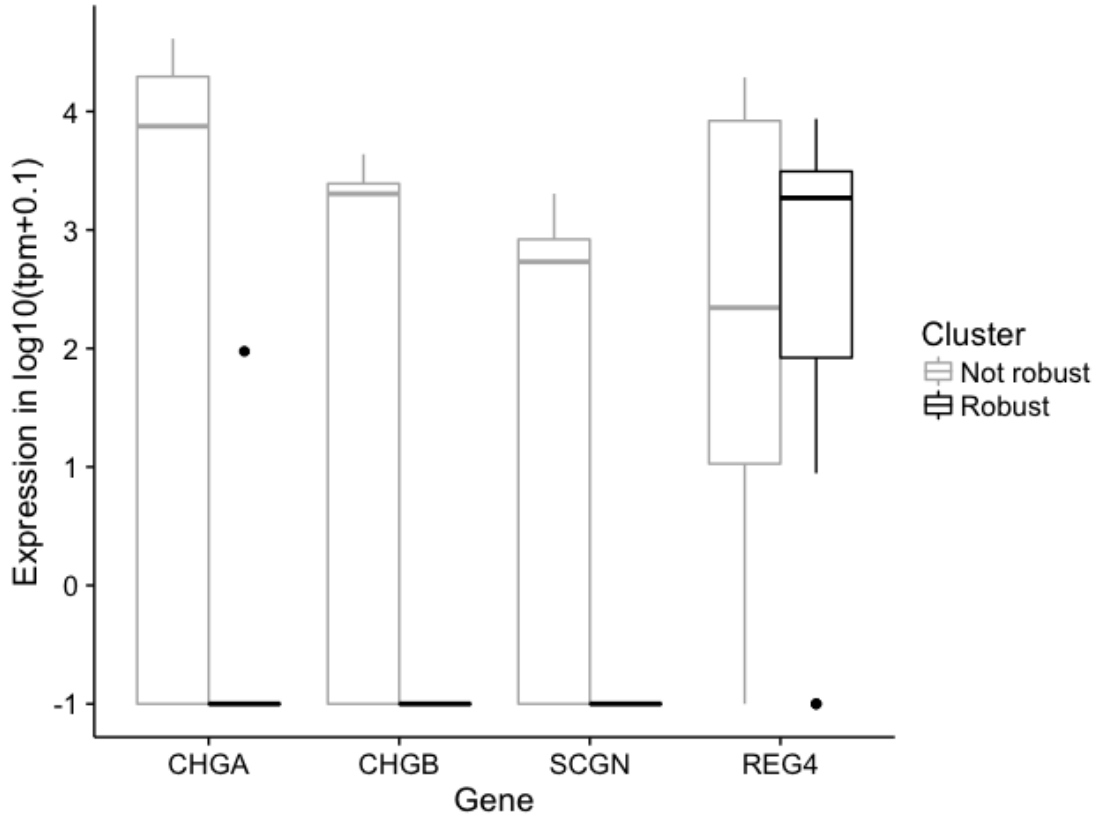


Figure A.1.31: Boxplots display the distribution of gene expression in log10(TPM+1) for known entero-endocrinocyte genes (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. The color of the boxplots denote whether the distribution consisted of cells in cluster 5 that were included in differential expression analysis (robust) or excluded (not robust).

A.1.6 Barrett's tissue cluster interpretation

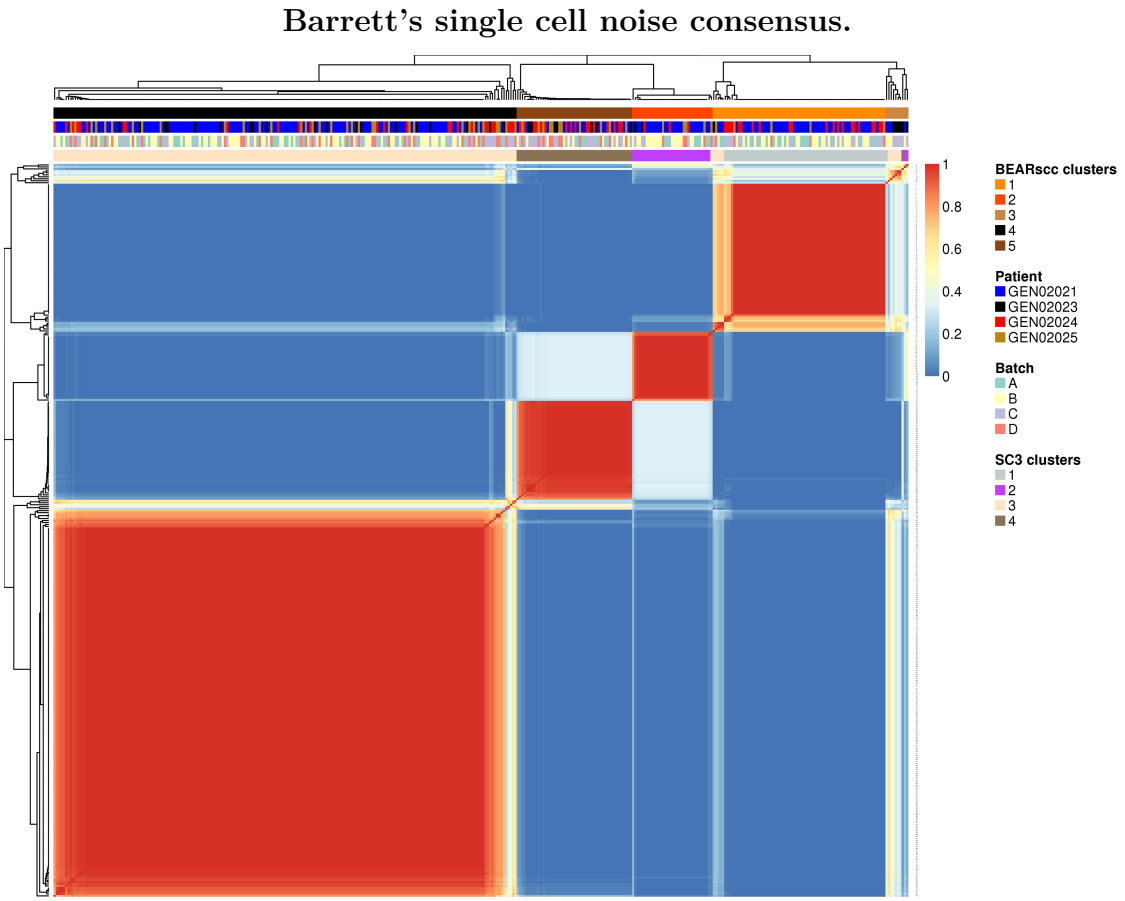


Figure A.1.32: The noise consensus of the Barrett's esophagus cells. Heatmap displays the noise consensus from SC3 clustering with $k = 4$ of 100 simulated technical replicates of cells isolated from endoscopic biopsies of the Barrett's lesion. Elements of the heatmap indicate the association frequencies of cells ordered by complete hierarchical clustering of the noise consensus. The 5 BEARscc clusters previously selected using the score metric distribution of clusters are annotated by color above the heatmap. Patient, batch and the initial SC3 clustering with $k = 4$ clusters of the observed data are also annotated by color above the heatmap.

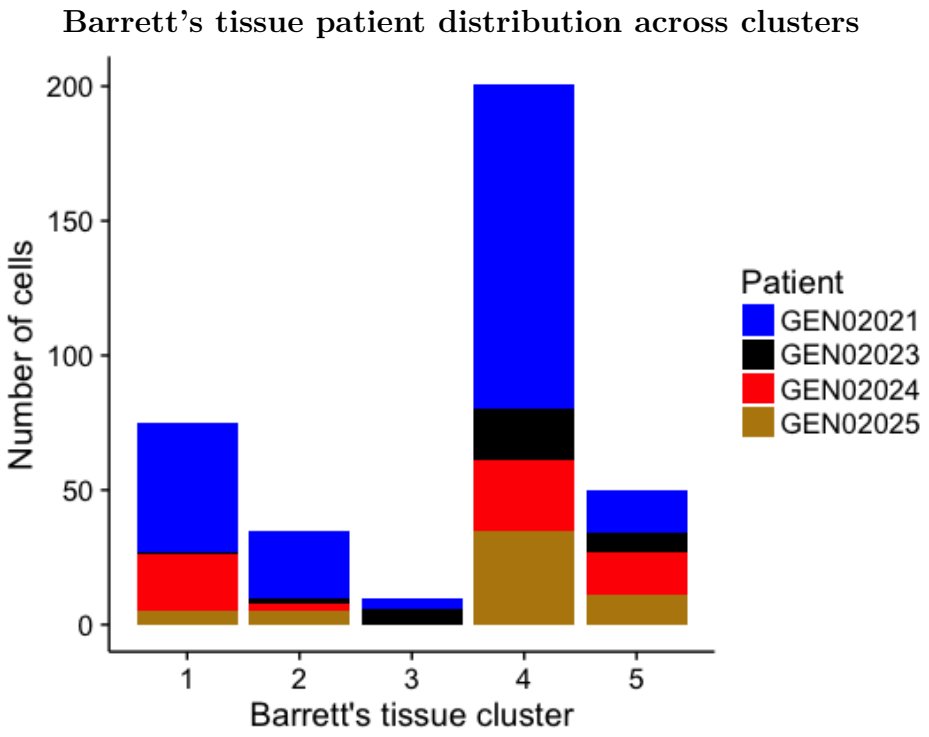


Figure A.1.33: The Barrett's esophagus clusters do not appear to be confounded by patient source. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

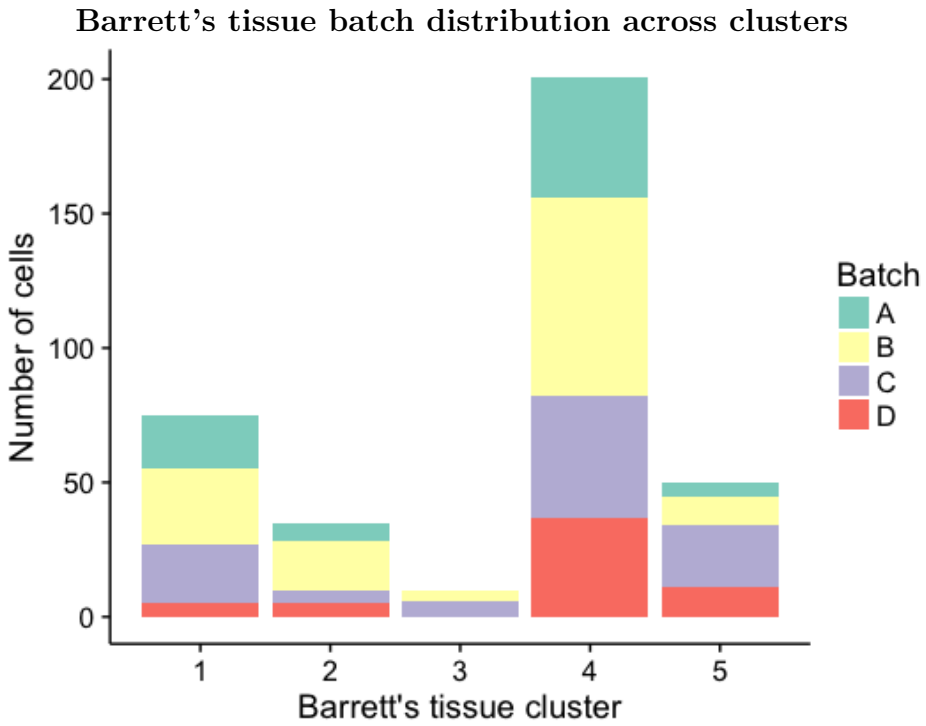


Figure A.1.34: The Barrett's esophagus clusters do not appear confounded by batch. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by batch from which the single cell samples were isolated.

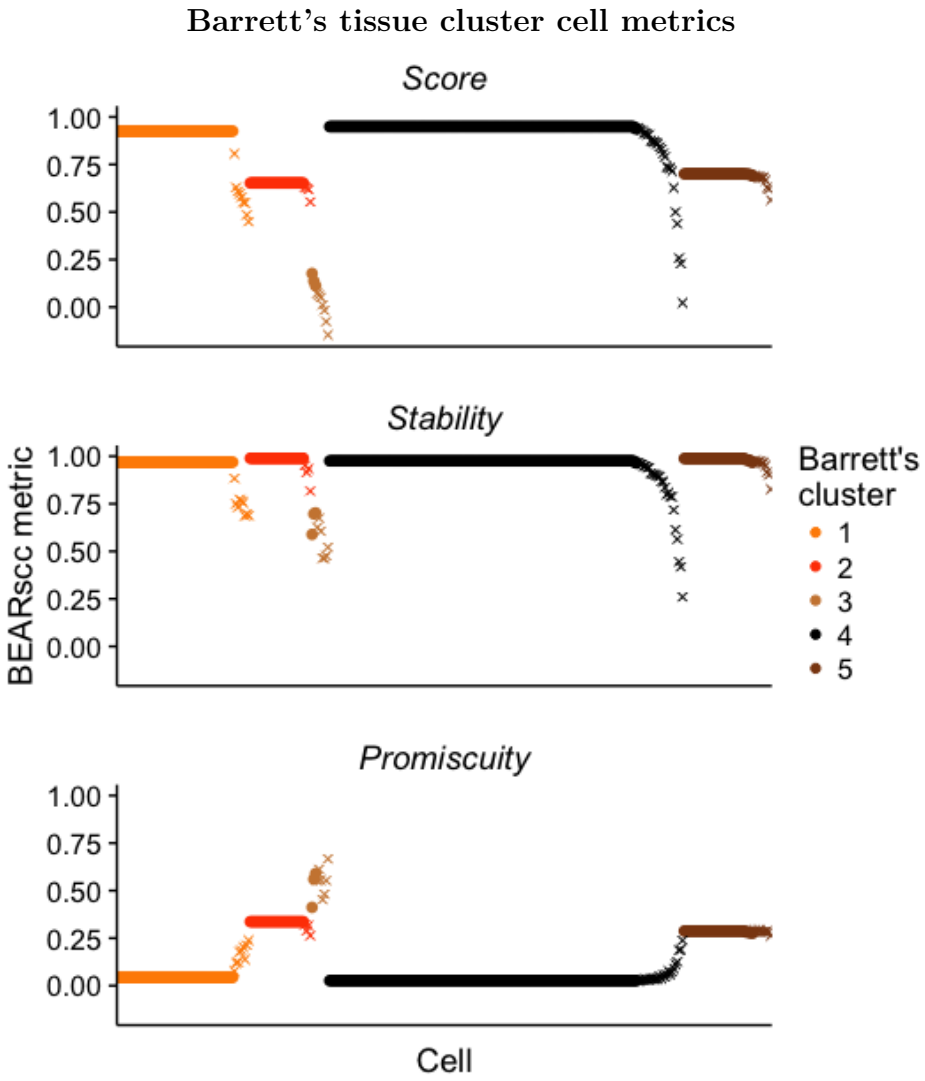


Figure A.1.35: Scatterplots display the BEARscc metrics (panels: score, stability, promiscuity) for each Barrett's cell (x-axis) ordered by the Barrett's tissue cluster to which it belongs and the BEARscc score metric. Barrett's single cell clusters are annotated by color. Inclusion for differential expression analysis is indicated by dots, while exclusion is indicated by crosses.

Brief explanation of interpretation Given the low score for cluster 3, and the fact that only 3 cells were utilized for differential expression analysis, it was not surprising that I did not observe differentially expressed genes in cluster 3. Among the differentially expressed genes in the other four BEARscc clusters, I found that several genes known to be important in Barrett's esophagus biology were present (Figure A.1.36). BEARscc clusters 1 and 2 in the Barrett's tissue was easily recognized as well-differentiated secretory cells. Cluster 1 exhibited high expression in mucin-2 (MUC2), trefoil factor 3 (TFF3), and the secrotry lineage transcription factor downstream of NOTCH (HES6). Together these observations suggested that cluster 1 was composed of secretory goblet-type cells. As I observed in the duodenal samples, two potential novel markers of goblet cell lineage, ITLN1 and SPINK4, were among the most significant genes differentially expressed in the goblet cell cluster 1. In cluster 2 I observed high expression of entero-endocrine cell markers: chromogranin-A (CHGA), chromogranin-B (CHGB), and secretogranin (SCGN). Therefore, I concluded that cluster 2 was principally composed of entero-endocrine cells.

In the other three BEARscc clusters of the Barrett's tissue, there were less obvious markers of well-differentiated cell types. As I observed in the undifferentiated duodenal cluster 3, I found that Barrett's cluster 4 was primarily characterized by a lack of expression of marker genes of well-differentiated cell types. Indeed, 91 of the 93 significantly differentially expressed genes in Barrett's cluster 4 were identified as significantly under-expressed compared to the rest of the Barrett's cells. Only two genes, a mitochondrial DNA gene product (MTRNR2L12) and an alcohol dehydrogenase (ADH1B), were identified as significantly over-expressed compared to the rest of the Barrett's tissue libraries.

Amongst differentially expressed genes, I observed that Barrett's cluster 5 exhibits a profile similar to that of Barrett's cluster 4 (Supplementary Figure A.1.37); however, cluster 5 is unique in that the cells within cluster 5 express genes associated with proliferation, e.g. cyclin-dependent kinase 1 (CDK1), cyclin B (CCNB1), and KI67 (MKI67) (Figure A.1.36). Interestingly, I noted that OLMF4

was significantly under-expressed in the well-differentiated Barrett's clusters 1 and 2, but highly expressed in the other three less well-differentiated clusters. This suggests that OLMF4 could be a marker for undifferentiated or stem-like genes in Barrett's esophagus, but further experimentation would be required to confirm or deny such a hypothesis.

While Barrett's cluster 3 had no significantly differentially expressed genes associated with it, I observed that amongst the curated marker genes, cells in cluster 3 were high expressers of the OLMF4, SPINK4 and HES6, but poor expressers of goblet cell marker MUC2. I previously postulated that SPINK4 could be an early marker of goblet cell differentiation and HES6 is an early transcriptional NOTCH-related signal in the intestine targeting cells for a secretory cell fate. Furthermore, in the noise consensus I observed that cells in cluster 3 associated strongly with the undifferentiated cells in cluster 4 across perturbations and a subset of the goblet-like well-differentiated cells in cluster 1. Taking these observations collectively, I hypothesized that cluster 3 consist of early secretory progenitors. However, further analysis is necessary to validate or refute this postulation.

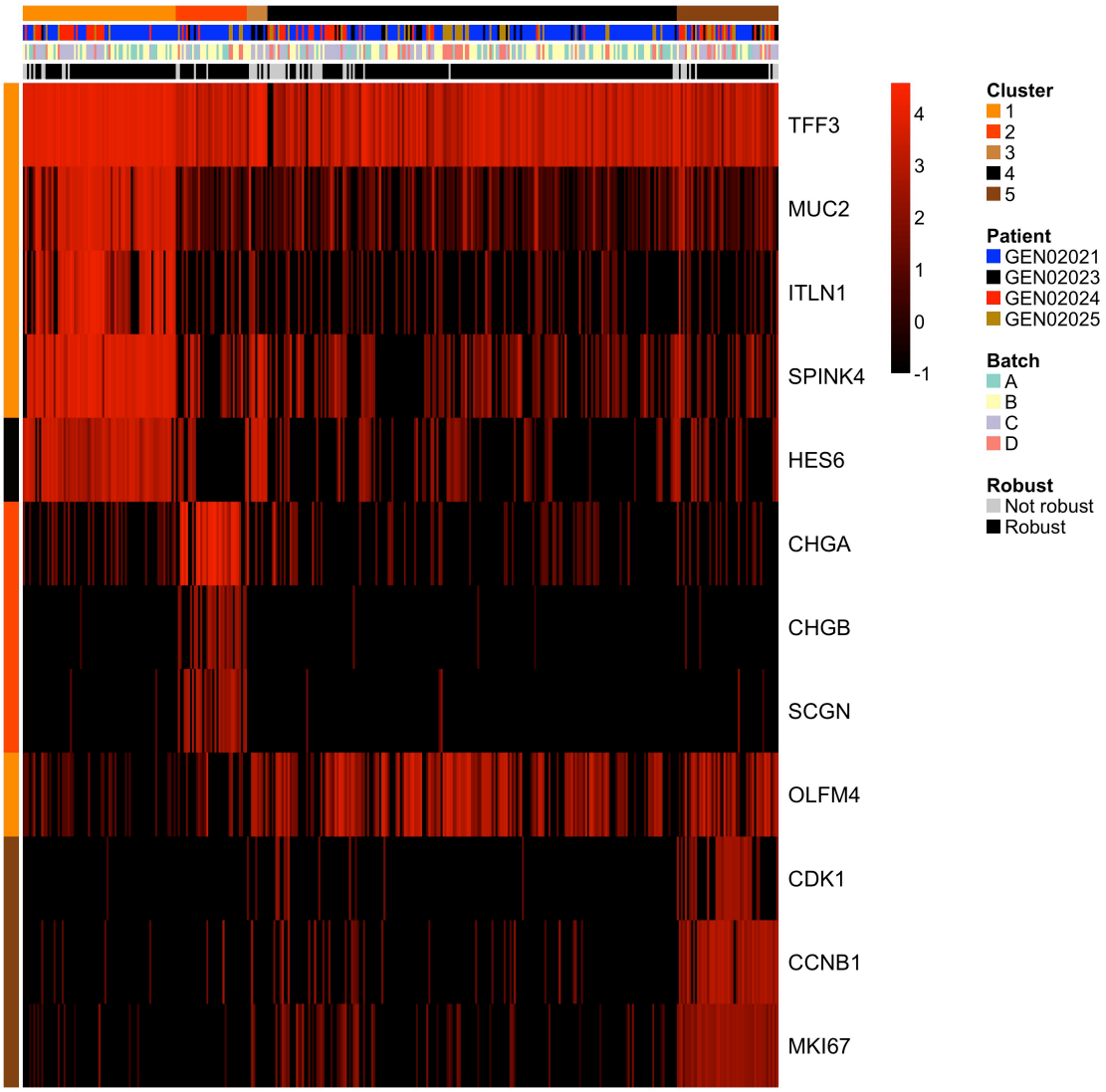


Figure A.1.36: Heatmap displays gene expression in Barrett's cells in log10(TPM+0.1) for select differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

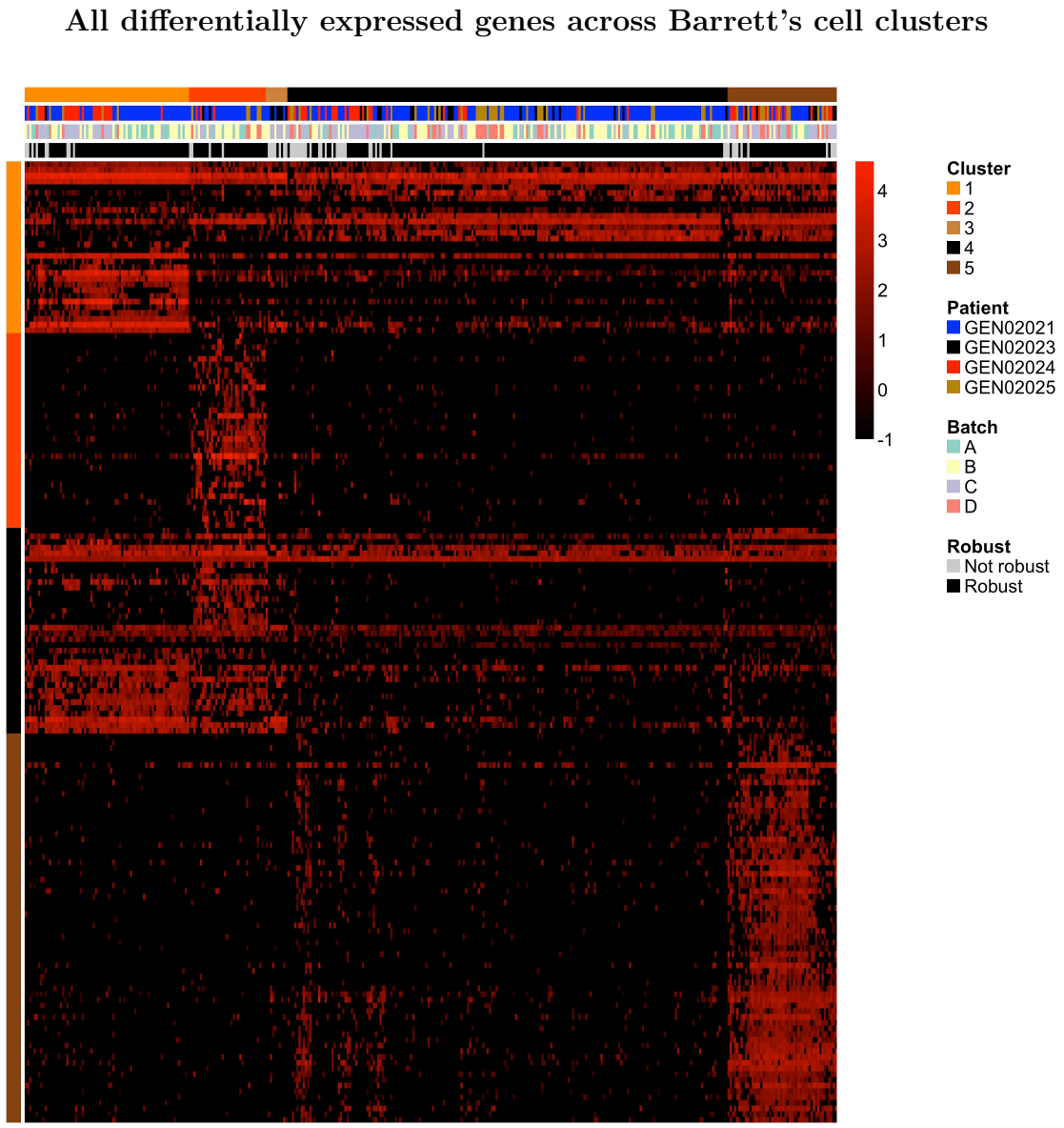


Figure A.1.37: Heatmap displays gene expression in Barrett's cells in log10(tpm+0.1) for differentially expressed genes. For each cell, the cluster, patient, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

Closer examination of Barrett's tissue cluster expression Because it has been suggested that Barrett's esophagus arises from a stem cell with properties similar to those of the intestine, I examined the expression of LGR5 in cells isolated from Barrett's tissue across the BEARscc Barrett's clusters (Figure A.1.38). In similar fashion to the undifferentiated duodenal cluster 3, the undifferentatiated Barrett's cluster 4 was rich in LGR5 positive cells. While BEARscc Barrett's clusters 1, 3, and 5 contained few LGR5 positive cells (3, 1, and 1, respectively), cluster 4 contained 24 LGR5 positive cells. Intriguingly, the entero-endocrine cluster contained 7 LGR5 positive cells, and the expression of LGR5 in these cells was comparable to the higher LGR5 expressing cells in cluster 4. Taken together this evidence suggests that Barrett's cluster 4, in like fashion to duodenal cluster 3, contains a gradient of undifferentiated cells and likely includes the least differentiated stem cells that populate and differentiate into the Barrett's gland.

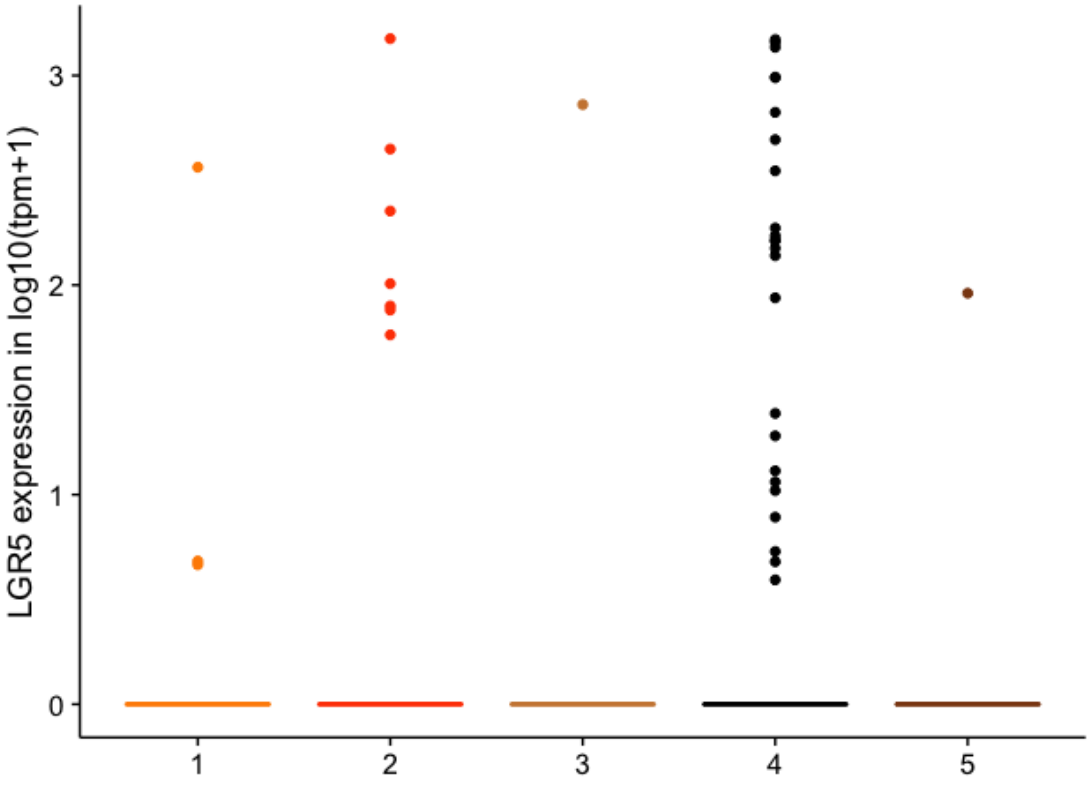


Figure A.1.38: Boxplots display the distribution of LGR5 expression in log10(TPM+1) by Barrett's BEARscc cluster (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. BEARscc Barrett's clusters are also coded by color.

A.1.7 All QC-filtered cell cluster interpretation

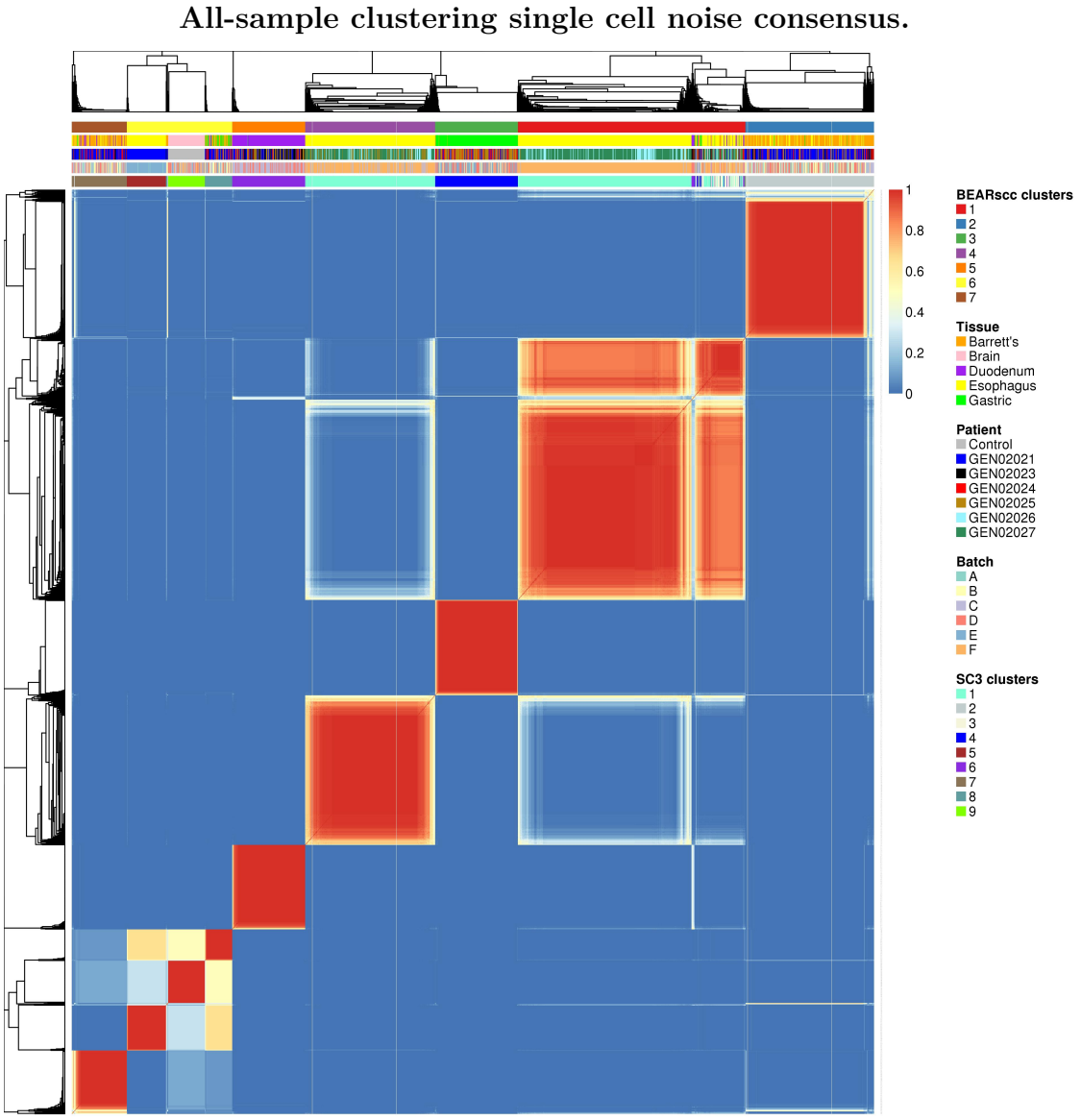


Figure A.1.39: Heatmap displays the seven robust clusters and the consensus of noisy counts matrices across tissues.

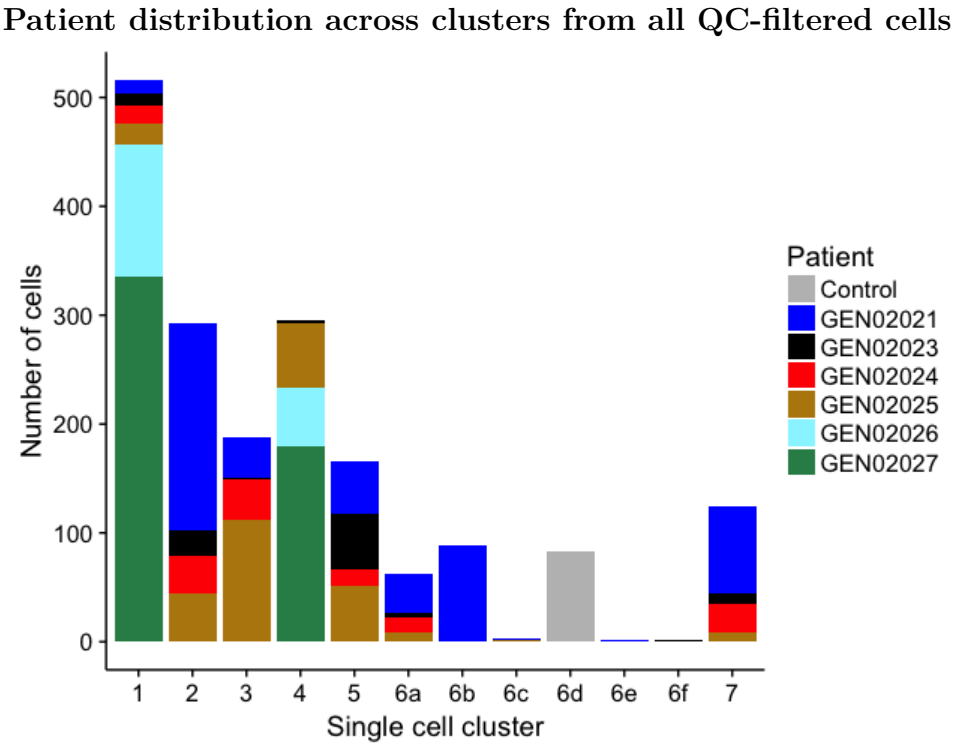


Figure A.1.40: The single cell clusters do not appear to be confounded by patient source. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by patient from which the single cell samples were isolated.

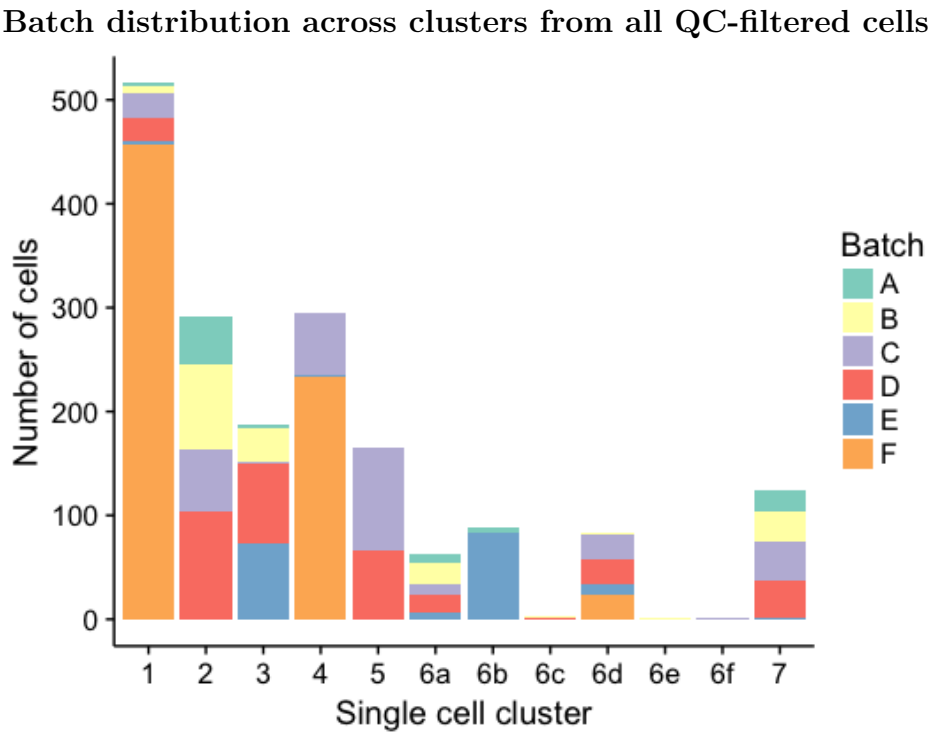


Figure A.1.41: The single cell clusters from all QC-filtered samples do not appear to be confounded by batch. Stacked barplot displays the number of cells (y-axis) in each cluster (x-axis), where bar is colored by batch from which the single cell samples were isolated.

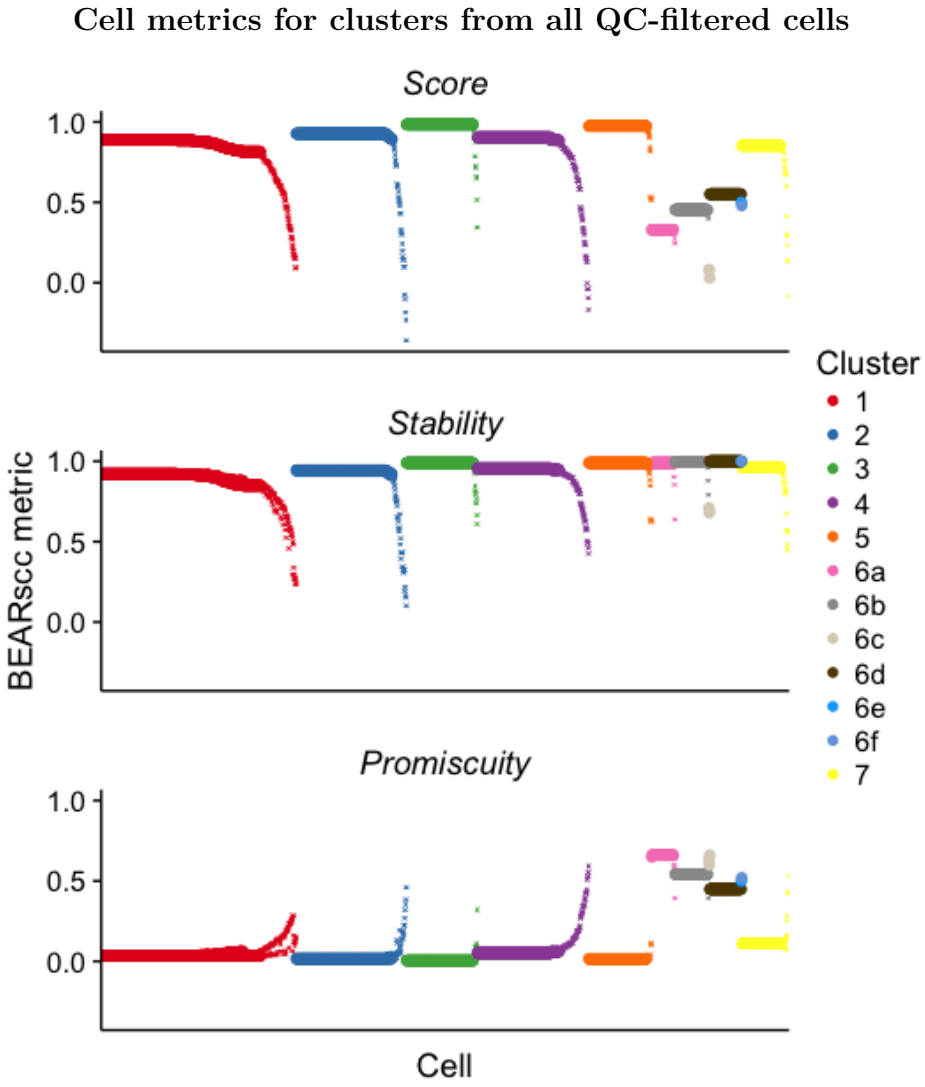


Figure A.1.42: Scatterplots display the BEARscc metrics (panels: score, stability, promiscuity) for each QC-filtered sample (x-axis) ordered by the cluster to which it belongs and by the BEARscc score metric. Single cell clusters are annotated by color. Inclusion for differential expression analysis is indicated by dots, while exclusion is indicated by crosses.

Brief explanation of interpretation In order to understand the biology behind the BEARscc clusters, I performed differential expression analysis as described for individual tissue(see Section 3.3.6) comparing each cluster to the rest of the single cells. As mentioned previously, clusters 6c, 6e, and 6f were too small to characterize using differential expression and were excluded from analysis. In the other 9 curated BEARscc clusters, I identified a number of significantly differentially expressed genes (Supplementary Figure A.1.44 and Supplementary Table A.13). The number of significantly over-expressed and under-expressed genes observed for each BEARscc cluster is listed in Table A.14. Interestingly, clusters 2 and 4 were dominated by under-expressed genes, while the majority of clusters were dominated by over-expressed genes. Among the differentially expressed genes in each cluster, I observed several genes with clear biological implications (Figure A.1.43).

All three BEARscc clusters consisting of primarily of esophageal cells (clusters 1, 4, and 6b) were all positive for squamous esophageal markers keratins 5 and 13 (KRT5 and KRT13) (Figure A.1.43). Cluster 6b was unique from the other esophagus dominated clusters in that I observed high expression of stress marker keratin-17 (KRT17), interferon-induced guanylate binding protein 5 (GBP5), a chemokine (CXCL9), and a squamous basal maker galectin-7 (LGALS7). Together these proteins indicate that cluster 6b may have both more basal and inflammatory response character than the other two clusters. Intriguingly, cells in cluster 4 did express LGALS7 and KRT17, but less frequently and at lower levels than cells in cluster 6 b. A small fraction of cells in cluster 1 did express LGALS7, but at several orders of magnitude less than cluster 4. Collectively, these observations seemed to suggest that clusters 1, 4, and 6b were all squamous esophageal tissue, where cells existed on a inflammatory stress or differentiation gradient from the most stressed / least differentiated cluster 6b to the middle cluster 4 and onto the least stressed / most differentiated cluster 1.

While it was not surprising to find that the three clusters composed mostly of esophageal tissue had strong squamous character, I wanted to understand the presence of the columnar cells in the well-differentiated squamous esophagus cluster

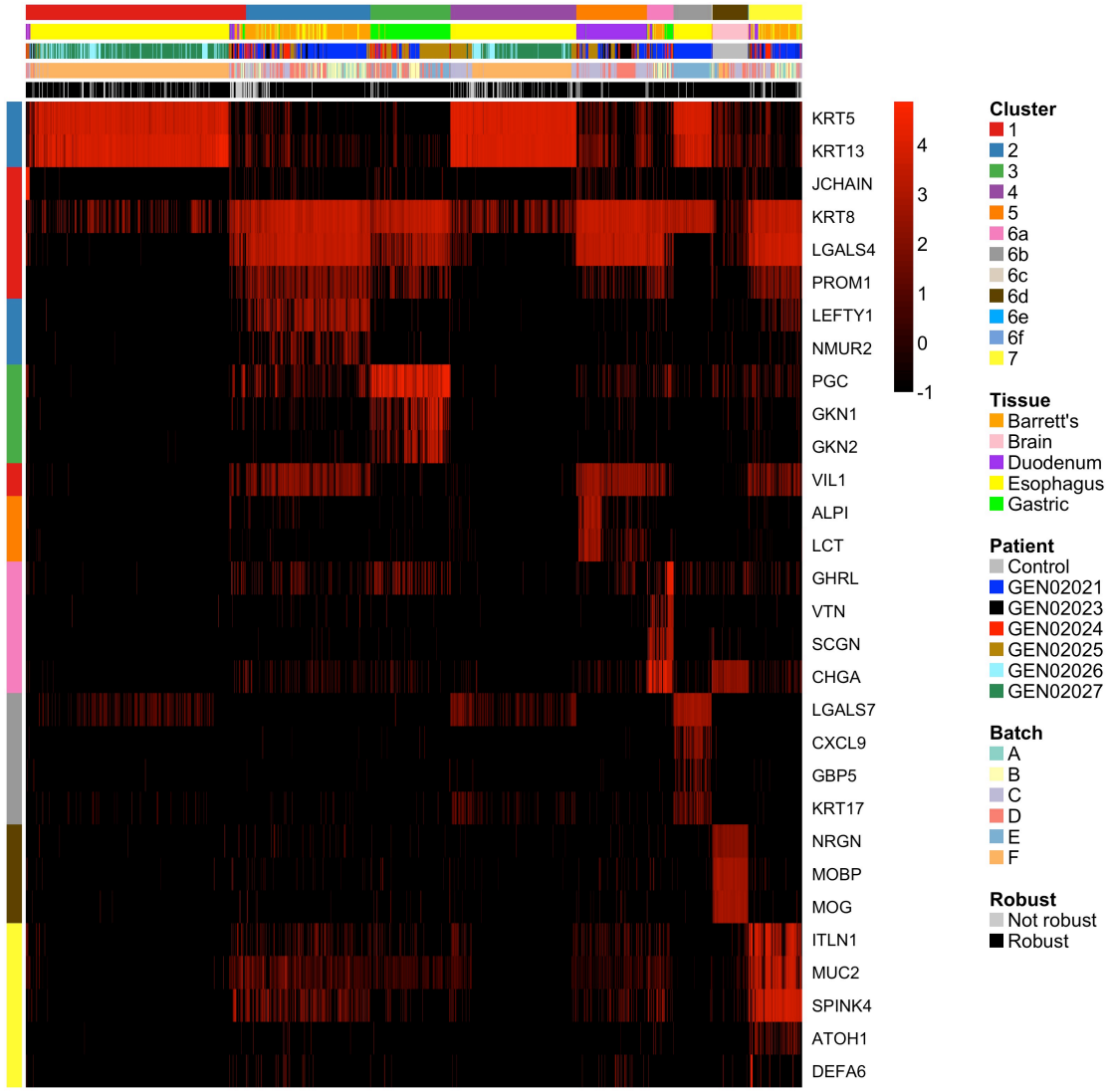


Figure A.1.43: Heatmap displays gene expression in single cells in log10(TPM+0.1) for select differentially expressed genes. For each cell, the cluster, patient, tissue, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

1. First, I noted that the cluster included a subset of intestinal and Barrett's cells expressing the J chain protein (JCHAIN), which suggests that these cells have non-epithelial character. In addition, the JCHAIN expressing columnar cells in cluster 1 also expressed heavy chain immunoglobulin molecules, IGHM and IGHA2

providing further evidence that a subset of the columnar cells in cluster 1 were not epithelial cells at all (Figure A.1.45 and Supplementary Figure A.13). However, these observations did not explain a number of other Barrett's, gastric, and duodenal cells in cluster 1 nor did it explain the reason for these non-epithelial cells clustering with well-differentiated squamous esophagus cells.

All differentially expressed genes across all QC-filtered cells

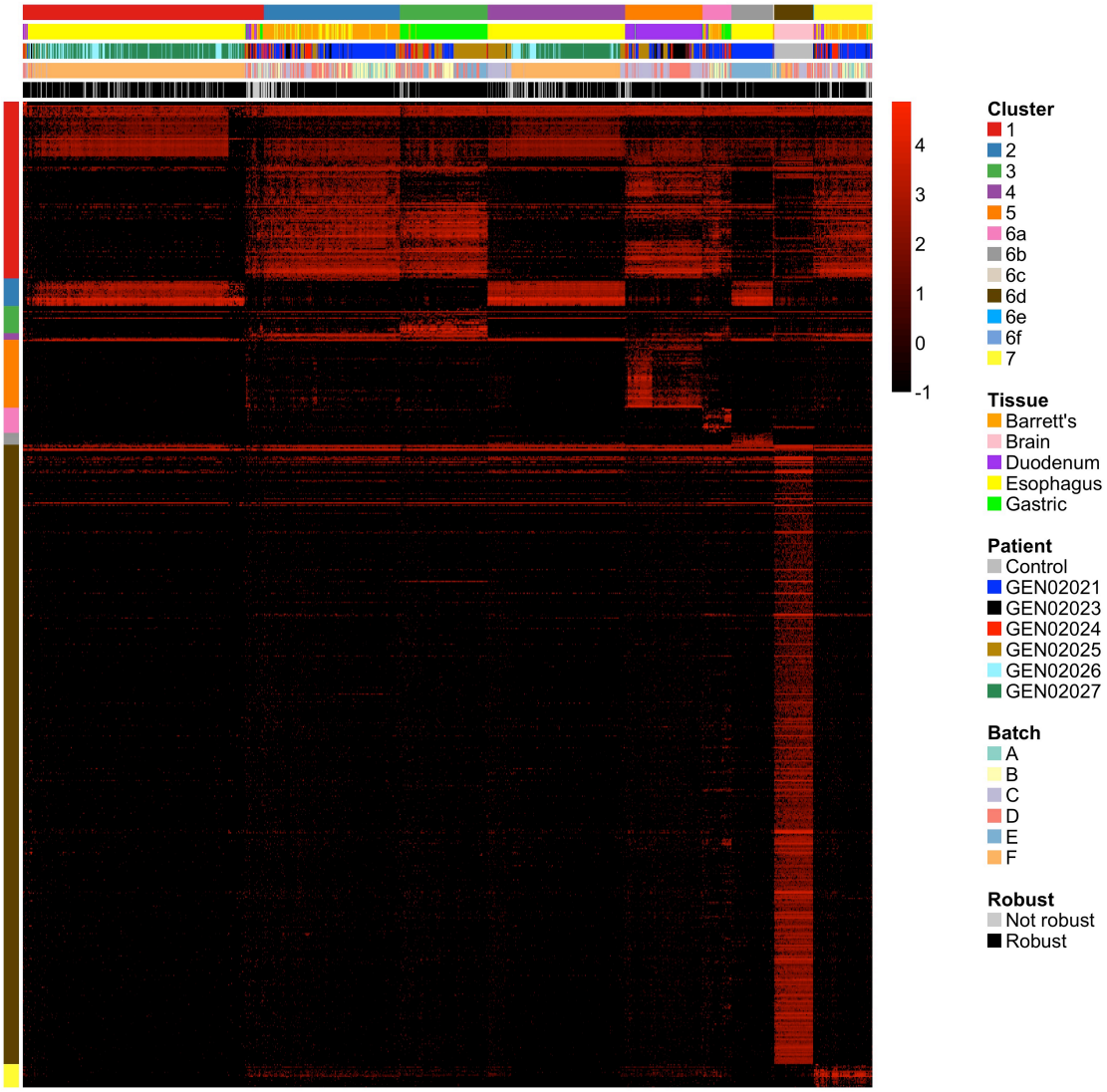


Figure A.1.44: Heatmap displays gene expression in single cells in log10(tpm+0.1) for differentially expressed genes. For each cell, the cluster, patient, tissue, and batch are annotated by color above the heatmap. Cells included in differential expression analysis due to low BEARscc score metrics are indicated in black and cell excluded in differential expression analysis are indicated by grey above the heatmap. For each gene, the cluster in which the gene was differentially expressed is indicated by color to the left of the heatmap.

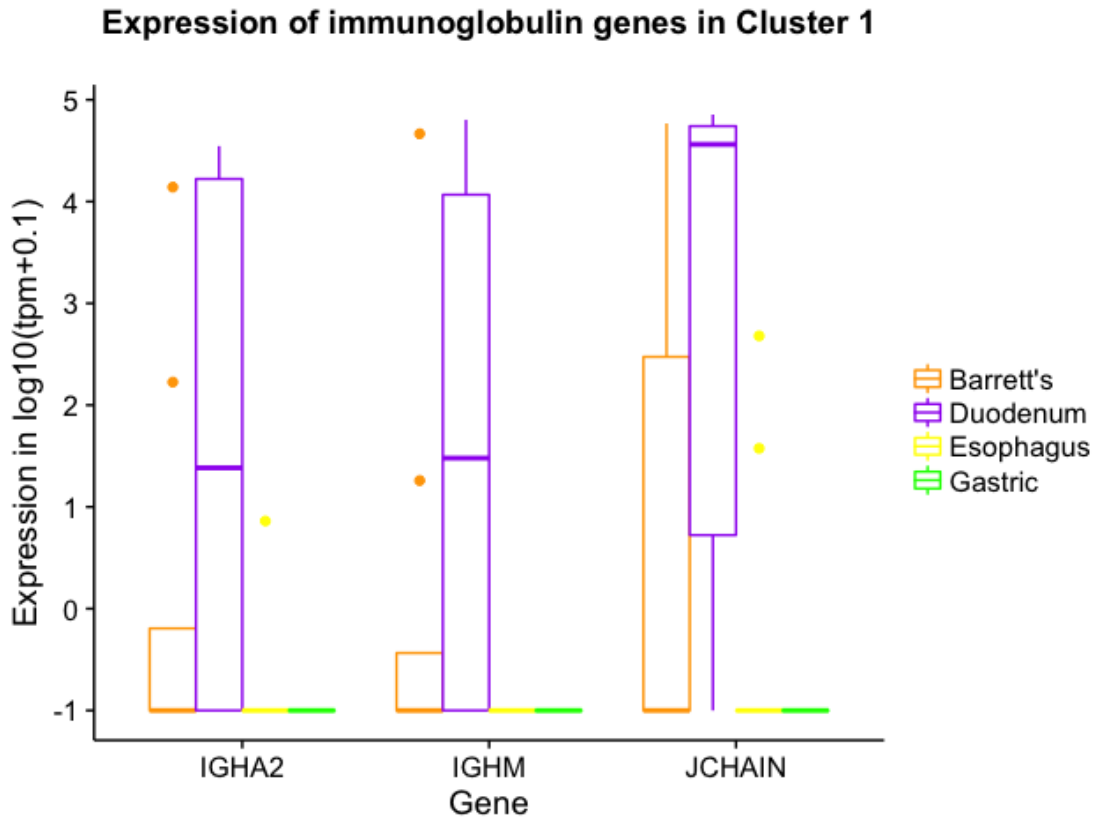


Figure A.1.45: Boxplots display the distribution of significantly over-expressed immunoglobulin gene expression in log10(TPM+0.1) (x-axis) for all-sample analysis cluster 1. Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. Tissues of origin are coded by color.

After examining the list of significantly over-expressed genes, I observed that 15 of those genes were ribosomal pseudogenes (Supplementary Table A.13). The high presence of ribosomal pseudogenes led us to suspect that the clustering was an artefact of a macroscopic feature of the libraries in question, so I examined the total gene expression of single cell libraries in the context of BEARscc cluster and tissue of origin (Figure A.1.46). As I hypothesized, BEARscc curated cluster 1 was populated with a large number of low expression cells compared to all other clusters and all columnar cells in cluster 1 with the exception of a single duodenal cell were among the low expression population. Therefore, I concluded that the columnar presence in cluster 1 consisted of a subgroup of non-epithelial cells and

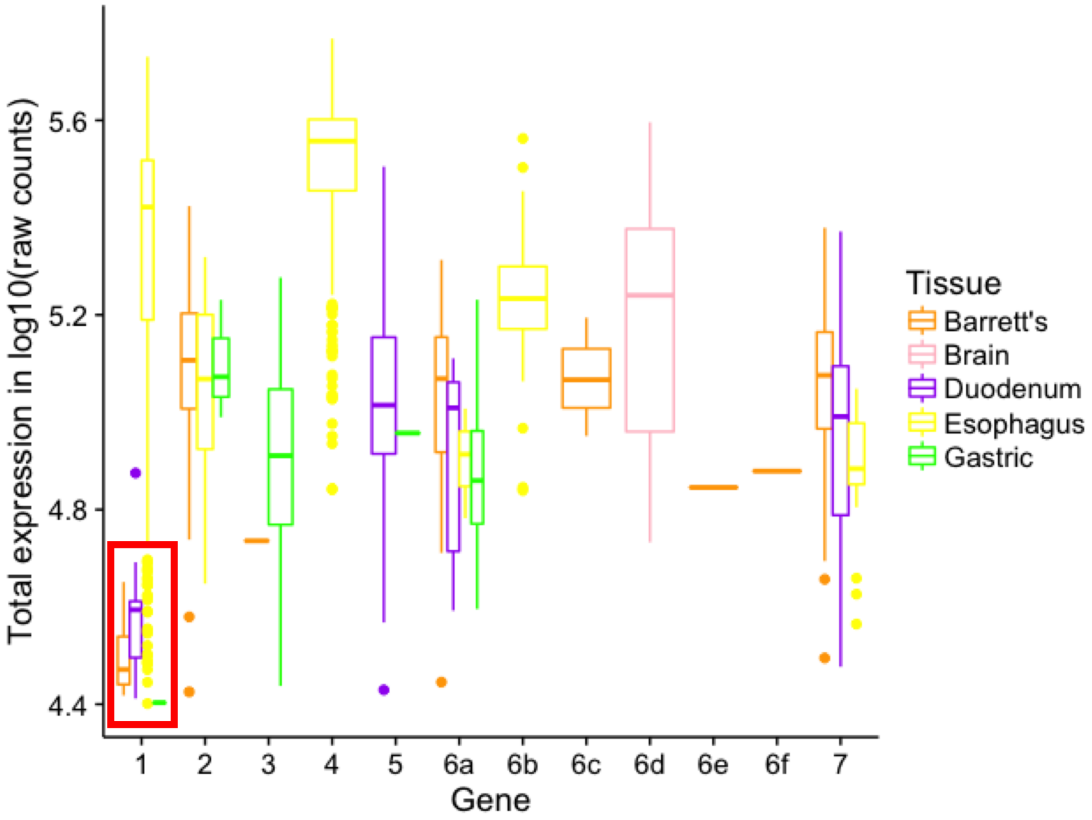


Figure A.1.46: Boxplots display the distribution of total gene expression in log10(raw counts+1) by single cell BEARscc cluster (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. Within each cluster, distribution are separated and colored by tissue of origin. The red square denotes the low expression samples in cluster 1 of columnar origin.

a subgroup of columnar cells that had clustered with the high number of low expression squamous esophageal cells present in our set of sample libraries. This conclusion was further supported by the observation that expression of differentially expressed genes strongly expressed in most cells of cluster 1 was visibly depressed for the columnar cells and a subgroup of esophageal cells in cluster 1 (Supplementary Figure A.1.44 - rightmost cells in cluster 1 annotated in red).

The predominantly cluster 3, the predominantly duodenal cluster 5, and cluster 6d entirely composed of brain control samples were more easily interpreted as compared to cluster 1. Among the significantly differentially expressed genes with over-expression in cluster 3 were the zymogen pepsinogen C (PGC), and the

gastric markers gastrokine-1 (GKN1) and gastrokin- 2 (GKN2). All three genes are markers for surface mucosa of gastric cells. Intriguingly, the single Barrett's cell in the cluster was positive for PGC, suggesting that the Barrett's cell had functional gastric character (Fig). Perhaps the cell was sampled from a well differentiated oxyntic Barrett's gland. In cluster 5 I observed the canonical intestine marker villin-1 (VIL1) expressed at high levels. Notably, the two clusters primarily consisting of cells from Barrett's tissue (clusters 2 and 7) also exhibit frequent and high VIL1 expression belying their intestinal nature. A subset of cells in cluster 5 expressed lactase (LCT) and intestine alkaline phosphatase (ALPI), known markers for absorptive enterocytes. Cluster 6d consisted only of whole tissue brain data and as expected was characterized by neuronal markers such as neurogranin (NRGN), myelin-associated oligodendrocyte basic protein (MOBP), and myelin oligodendrocyte glycoprotein (MOG).

Cluster 6b contained cells from several different tissues and associated with the brain control cluster 6d with high promiscuity. The differentially expressed high expressing genes unique to cluster 6b consisted of clear markers of entero-endocrine cells including vitronectin (VTN), secretogranin (SCGN), and chromogranin A (CHGA) (Figure A.1.43). Unique to the gastric members of the cluster was the expression of ghrelin and obestatin prepropeptide (GHRL), which is more commonly known as the hunger hormone. It was interesting to observe that enteroendocrine cells from different epithelia including the Barrett's lesion were functionally closer than with other cell types in their respective tissues of origin. I suspected that close relationship between neuronal and entero-endocrine cells was the reason for the high association with brain control cluster 6d.

Strong expression of mucin 2 (MUC2) suggests that cluster 7 consists of well-differentiated goblet-like cells. As with the entero-endocrinocytes, I was intrigued by the fact that secretory goblet cells from the Barrett's lesion, proximal squamous tissue, and the duodenum were functionally closer than with other sub-types sampled from the same respective tissue. While MUC2 expression was frequent and strong, SPINK4 was a better predictor for membership in the goblet-like cluster 7 regardless

of a cell's tissue of origin. SPINK4 has not been described as a goblet cell marker before, and I resolved to explore this observation further. ITLN1 was another gene that has yet to be described as a marker for goblet cells that was significantly differentially expressed with consistent high expression in cluster 7. In section 3.12.4, I describe how we examined these putative goblet cell marker genes more closely and Dr. Michael White performed immunofluorescence analysis to validate these results.

Cluster 2 was perhaps the most perplexing, and thus the most interesting cluster result in the BEARscc analysis of all QC-filtered cells. The cluster shared high expression of the columnar marker keratin-8 (Wang *et al.* 2009) with the gastric surface mucosa cluster 3, the intestinal cluster 5, and the well-differentiated entero-endocrine and goblet clusters (clusters 6a and 7, respectively) (Figure A.1.43). The galactose binding lectin galectin-4 (LGALS4) is an extra-cellular protein characterized as regulating lipid rafts in microvilli structures of columnar tissue (Cao & Guo 2016). I observed frequent LGALS4 expression in cluster 2 as well as the goblet cell cluster 7, the intestinal cluster 5, and to a much lesser extent the gastric surface mucosa cluster 3. Notably, LGALS4 expression was absent in the entero-endocrinocytes (cluster 6a) of all three columnar tissue types. The predominantly Barrett's and normal esophagus cluster 2 also exhibited frequent and high VIL1, a feature only found in the differentiated goblet cell cluster 7 and the intestinal cluster 5.

The expression pattern of significant differentially expressed genes in cluster 2 most closely resembled that of the intestinal cluster 5 (Figure A.1.44), with the notable exception of the absence of genes expressed in the absorptive enterocytes of intestinal cluster 5 (Figure A.1.44 - orange annotated genes of left-most cells of orange cluster 5). Indeed, all three clusters with high numbers of cells from Barrett's tissue did not contain any cells with obvious absorptive enterocyte character; from these results Barrett's appears to be purely composed of secretory lineages. I recalled that in the individual tissue clustering of the intestine and Barrett's esophagus, clusters D3 and B4 exhibited properties that suggested an undifferentiated and even stem cell phenotype including the presence of numerous LGR5 positive cells

and the absence of differentiation marker genes. In the all-sample clustering, cluster 2 is similarly bereft of well-differentiated marker genes (Table A.14) with only two genes (LEFTY1, NMUR2) identified as significantly overexpressed compared to the rest of the single cell samples. LEFTY1 is a developmental gene important in early embryogenesis cell fate and NMUR2 is a neural receptor with homology to rhodopsin, which was especially compelling given previous work demonstrating the potential for retinoic acid to transform squamous epithelia into a columnar phenotype. Given these observations, I thought it would be prudent to examine the expression of LGR5 in the context of the all-sample BEARscc clusters (Figure A.1.47). I observed a handful of LGR5 positive cells in the gastric cluster and in the duodenal clusters, which contained most of those cell types. However, the richest collection of LGR5 cells in a predominantly Barrett's cluster was found in cluster 2. Intriguingly a number of the LGR5 positive cells in the cluster were from the proximal normal esophagus. Taken together, these observations suggested that cluster 2 consisted of undifferentiated Barrett's cells and cells from the esophagus of similar functional character. I explore the presence of cells of proximal normal squamous esophageal cells further in Section 3.12.2.

Visualizing the relationship between patient, tissue, and cluster Upon interpreting the functional character of each BEARscc cell clustering group, I directly visualized the relationship between the interpreted clusters and the patient and tissue origin of cells (Supplementary Figure A.1.49). Comfortingly, the brain tissue positive control formed a single cluster, 6d. Cells from the duodenum predominantly clustered in the intestinal cluster 5; however, a small number of cells clustered in the well-differentiated entero-endocrinocytes (cluster 6a), goblet cells (cluster 7), and the low expression / non-epithelial cells cluster in the largest squamous esophagus cluster (cluster 1). The bulk of cells derived from gastric tissue clustered in the gastric surface mucosa cluster (cluster 3) and smaller sizeable number clustered with the entero-endocrinocytes of other tissue types (cluster 6a). As with the duodenum, a handful of low expression cells clustered in the largest squamous esophagus cluster

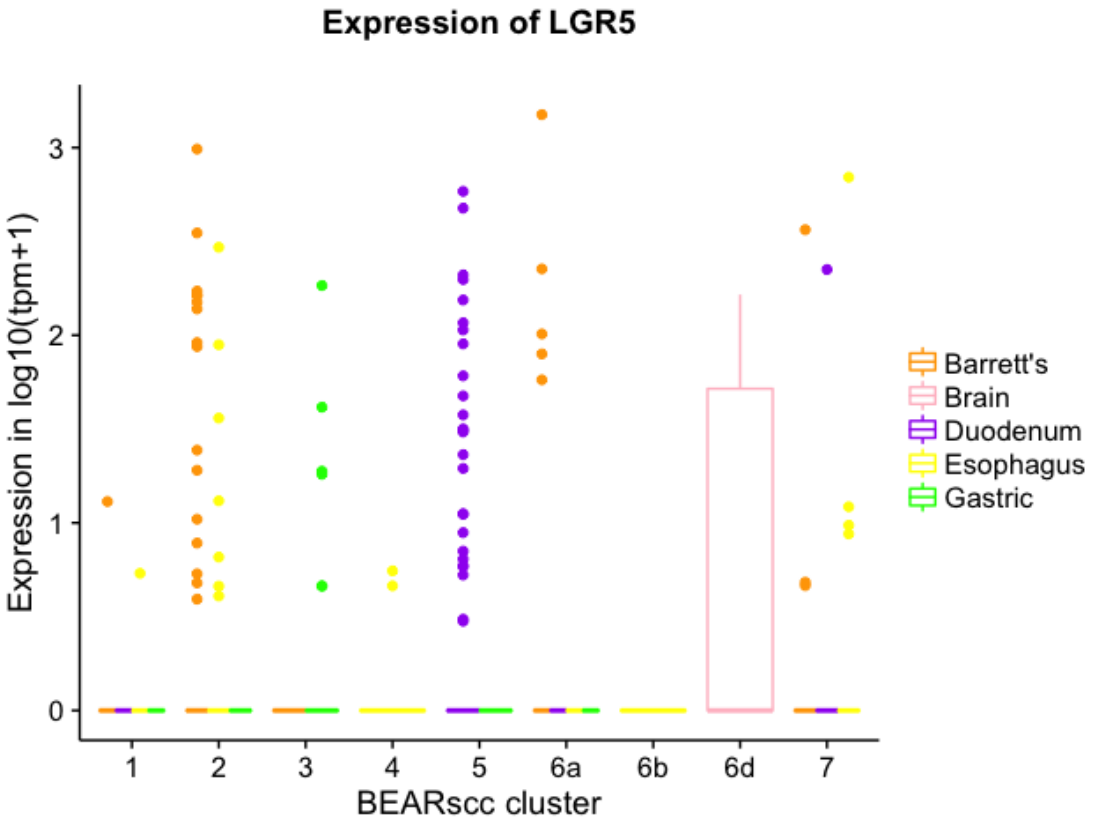


Figure A.1.47: Boxplots display the distribution of LGR5 expression in log10(TPM+1) by BEARscc cluster (x-axis). Within boxplots, the bold bar indicates median and the upper and lower hinges indicate the upper and lower quartiles respectively. Whiskers terminate at 1.5 times the interquartile range above and below the upper and lower quartiles. Values outside of the range illustrated by whiskers are displayed as dots. Distributions for each cluster are divided and indicated by tissue of origin.

(cluster 1). Intriguingly, one or two cells clustered with intestinal, Barrett's type, and goblet-like groups (clusters 5, 2, and 7, respectively), suggesting these cells were a segment of gastric tissue with intestinal metaplasia. However, there were to few cells to explore this hypothesis further. The cells form Barrett's esophagus formed the 3 small subclusters of cluster 6 (cluster 6c, 6e, 6f) for which I did not have sufficient power to characterize. As with the other columnar tissue types, a small number of low expressing Barrett's cells clustered with the largest squamous esophagus cluster (cluster 1). The majority of cells from the Barrett's lesion formed the relatively undifferentiated Barrett's type cluster (cluster 2). The rest of the cells isolated from Barrett's tissue clustered with other entero-endocrinocytes (cluster

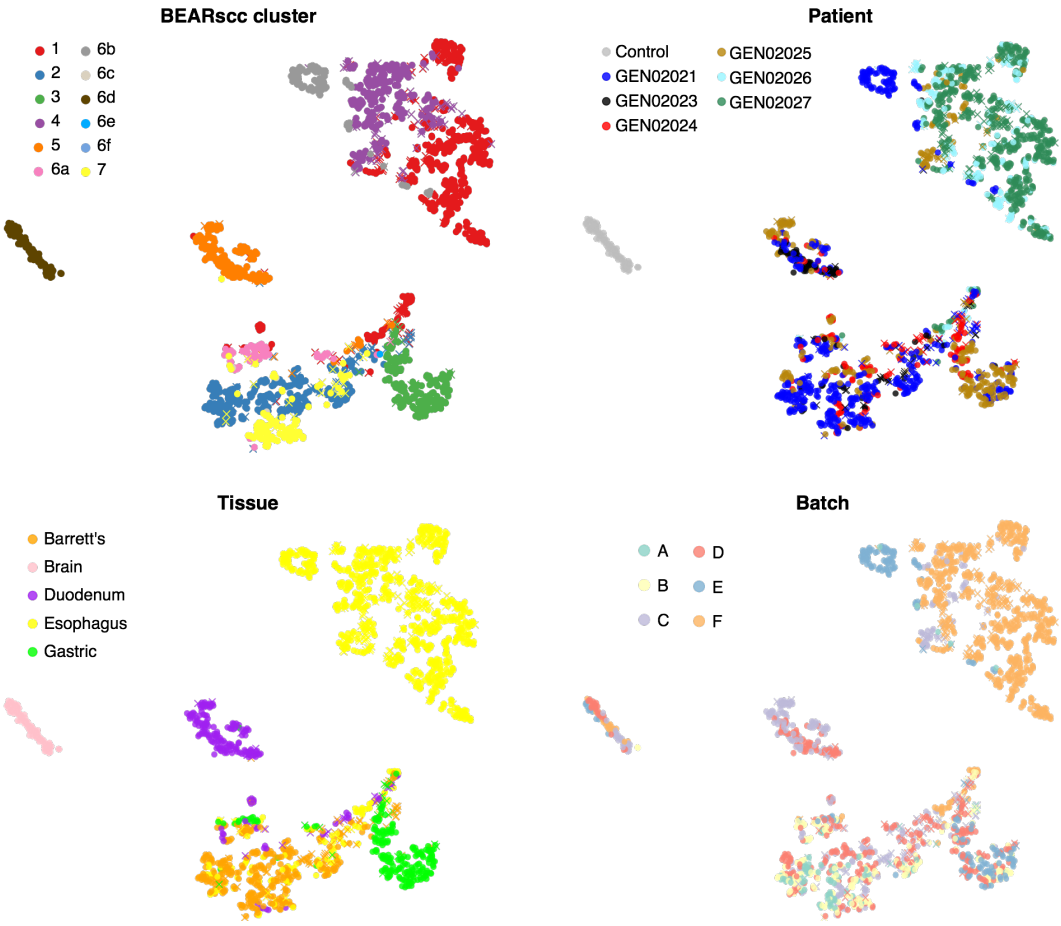


Figure A.1.48: Two-dimensional visualizations of all QC-filtered single cells using t-stochastic neighbor embedding (t-SNE) highlighting sample metadata characteristics. Single cells are colored by the assigned BEARscc cluster (upper left), the patient of origin (upper right), the tissue of origin (bottom left), or the experimental batch (bottom right).

6a) and goblet cells (cluster 7) from the other tissue types. Unsurprisingly, the cells isolated form the squamous esophagus were by and large seperate from the columnar cells forming the three squamous esophageal clusters (clusters 1,4, and 6b). Intriguingly, a small number of esophageal cells clustered with goblet (cluster 7) and entero-endocrine (cluster 6a) cells. Perhaps the most interesting observation was the relatively large number of esophageal cells that clustered with Barrett's cells in the undifferentiated Barrett's type cluster.

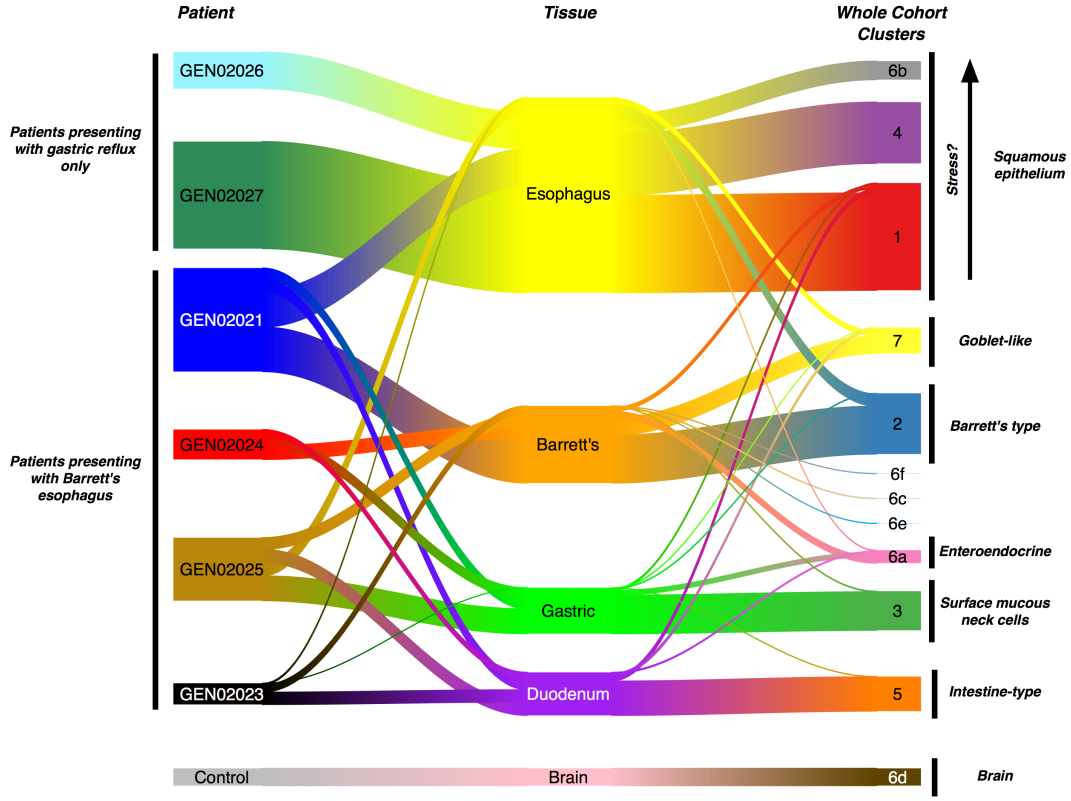
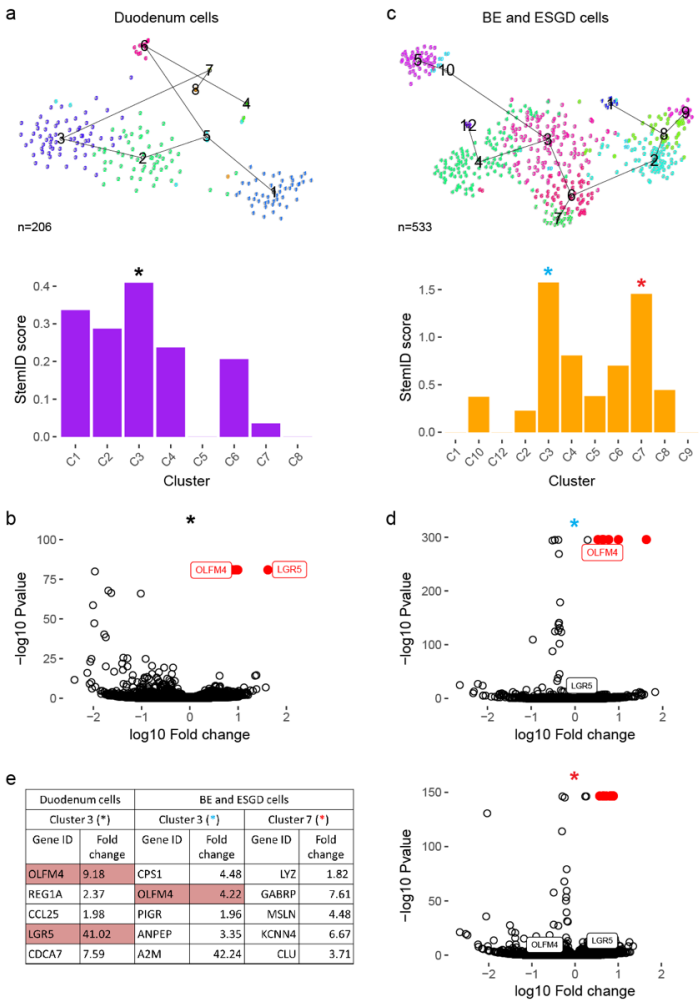


Figure A.1.49: Sankey diagram displays relationship between the patient origin of the sample (left), the tissue origin (middle) of the sample and the BEARscc clusters assigned to the sample (right). Nodes represent patient (left), tissue (middle), or BEARscc cluster (right), where nodes are indicated by color and node height represents the number of single cell samples in the respective calssification. Likewise edge size indicates the number of cells with shared classification between the two nodes to which it connects. Patient presentation is annotated in bold italics to the left, and biological interpretation of BEARscc clusters is indicated in bold italics to the right.

A.1.8 Further analysis of all QC-filtered cells



Duodenum cells		BE and ESGD cells			
Cluster 3 (*)		Cluster 3 (*)		Cluster 7 (*)	
Gene ID	Fold change	Gene ID	Fold change	Gene ID	Fold change
OLFM4	9.18	CPS1	4.48	LYZ	1.82
REG1A	2.37	OLFM4	4.22	GABRP	7.61
CCL25	1.98	PIGR	1.96	MSLN	4.48
LGR5	41.02	ANPEP	3.35	KCNN4	6.67
CDCA7	7.59	A2M	42.24	CLU	3.71

Figure A.1.50: StemID applied to human duodenum, Barrett's esophagus, and esophageal submucosal glandular duct cells. **a**, (Top) Plot showing entropy and links between RaceID2 clusters (numbered) as computed by StemID (for more information see manuscript (Grün *et al.* 2016b)) applied to duodenal cells. (Bottom) StemID scores across all clusters computed for duodenum cells. Score are calculated from multiplicaiton of the entropy (spread from the cluster mean) and the number of cluster links arising from a given cluster. Highest scoring cluster is indicated by a black asterisk. **b**, Volcano plot of genes from the highest scoring cluster from duodenum cells (C3). Points coloured red indicate the most significant genes with a fold change greater than 2. Selected genes of interest are labelled. **c**, As in **a** but applied to Barrett's esophagus and esophageal submucosal glandular duct cells with the highest and second highest scoring clusters indicated with blue and red asterisks, respectively. **d**, as in **b** but applied to Barrett's esophagus and esophageal submucosal glandular duct cells in the highest two scoring clusters, labelled with blue and red asterisks, respectively (top and bottom), **e**, A table of the top five most significantly differentially expressed genes in the asterisked clusters from **a** and **c** with fold change values (against all other clusters) shown. Genes discussed in chapter text are highlighted. All figures and legend are the work of Dr. Richard Owen performing independent analysis with the StemID algorithm.

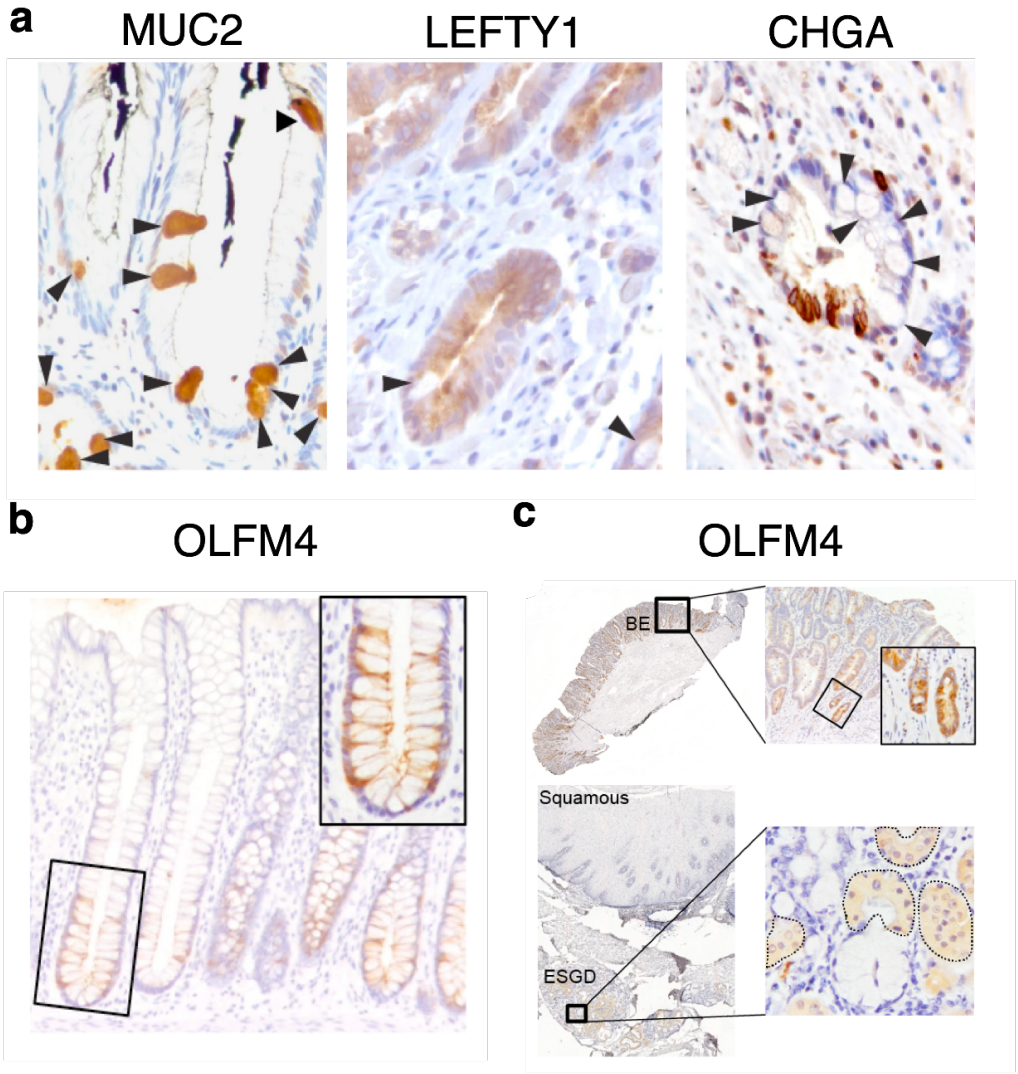


Figure A.1.51: **a**, Immunohistochemical staining of MUC2, LEFTY1, CHGA in adjacent sections of Barrett's esophagus resection specimens. Black arrows indicate goblet cells on all sections (positively stained for MUC2; negative for LEFTY1 and CHGA). **b**, Immunohistochemical staining of OLFM4 in human colon (close-up of base of crypt inset). **c**, Immunohistochemical staining of OLFM4 in Barrett's esophagus (top), and esophageal submucosal glandular ducts beneath squamous epithelium (bottom), with enlargements. Black dotted lines (lower right) indicate positively stained areas of OLFM4 in esophageal submucosal glandular ducts. Immunohistochemistry and legend are the work of Dr. Michael White.

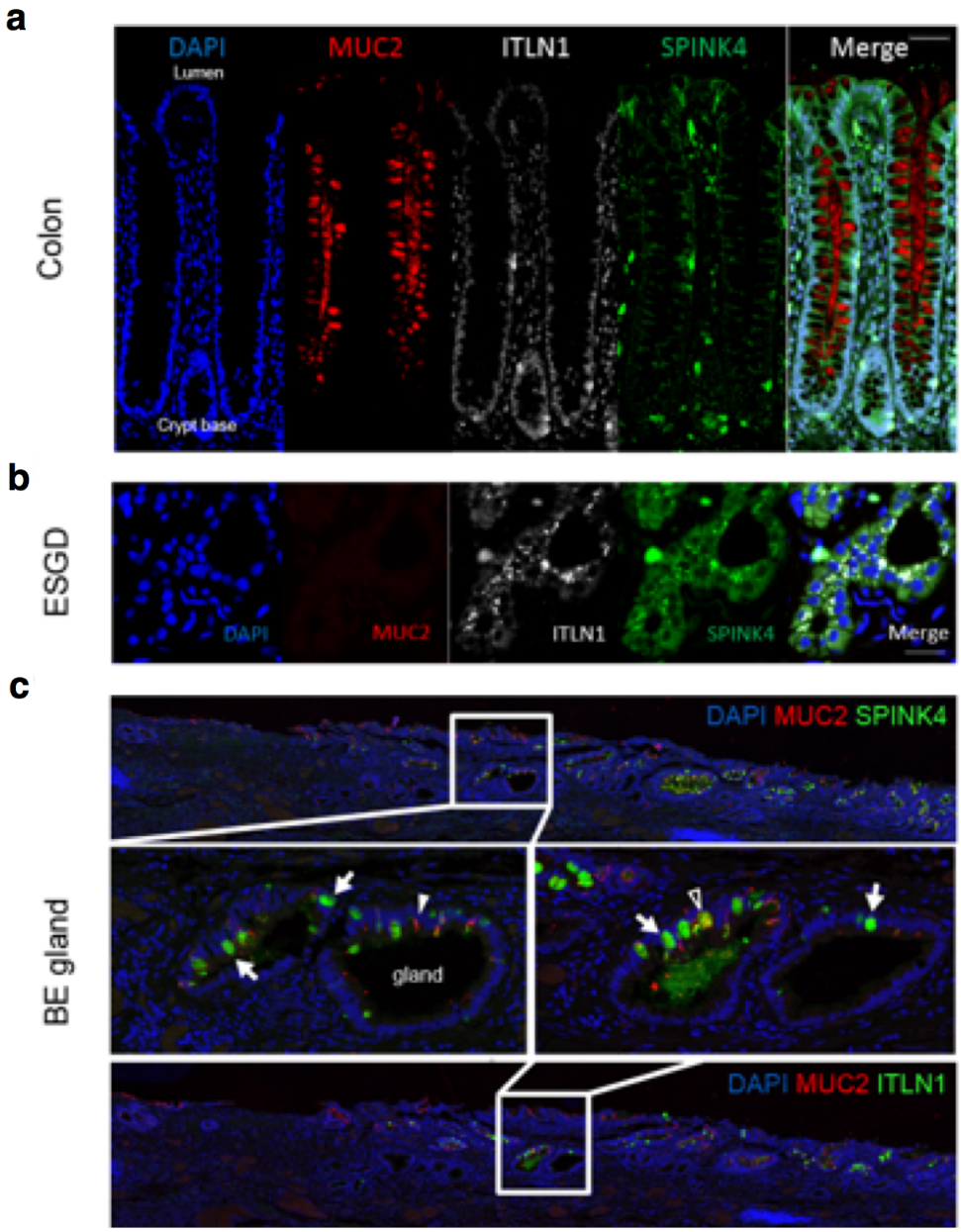


Figure A.1.52: **a**, Representative immunofluorescence staining of human colon crypts for MUC2 (red), ITLN1 (white) and SPINK4 (green); nuclei (DAPI) in blue. Scale bar (top right) indicates μm. **b**, Immunofluorescence staining of a representative esophageal submucosal glandular duct with MUC2 (red), ITLN1 (white) and SPINK4 (green). Scale bar (bottom right) indicates 50μm. **c**, Immunofluorescence staining of representative Barrett's esophagus sections with MUC2 (red) and SPINK4 (green) in the top image, and MUC2(red) and ITLN1 (green) in the lower image panel. Enlarged images of both sections are shown in the center. White tailed arrows show SPINK4 or ITLN1 expressing cells, white filled arrowhead shows a MUC2 expressing cell, white unfilled arrowhead indicates a cell co-staining for MUC2 and ITLN1. Immunofluorescence and legend are the work of Dr. Michael White.

A.1.9 Single cell pathogen detection

Detected pathogens in single cell samples from patients that presented with Barrett's esophagus.

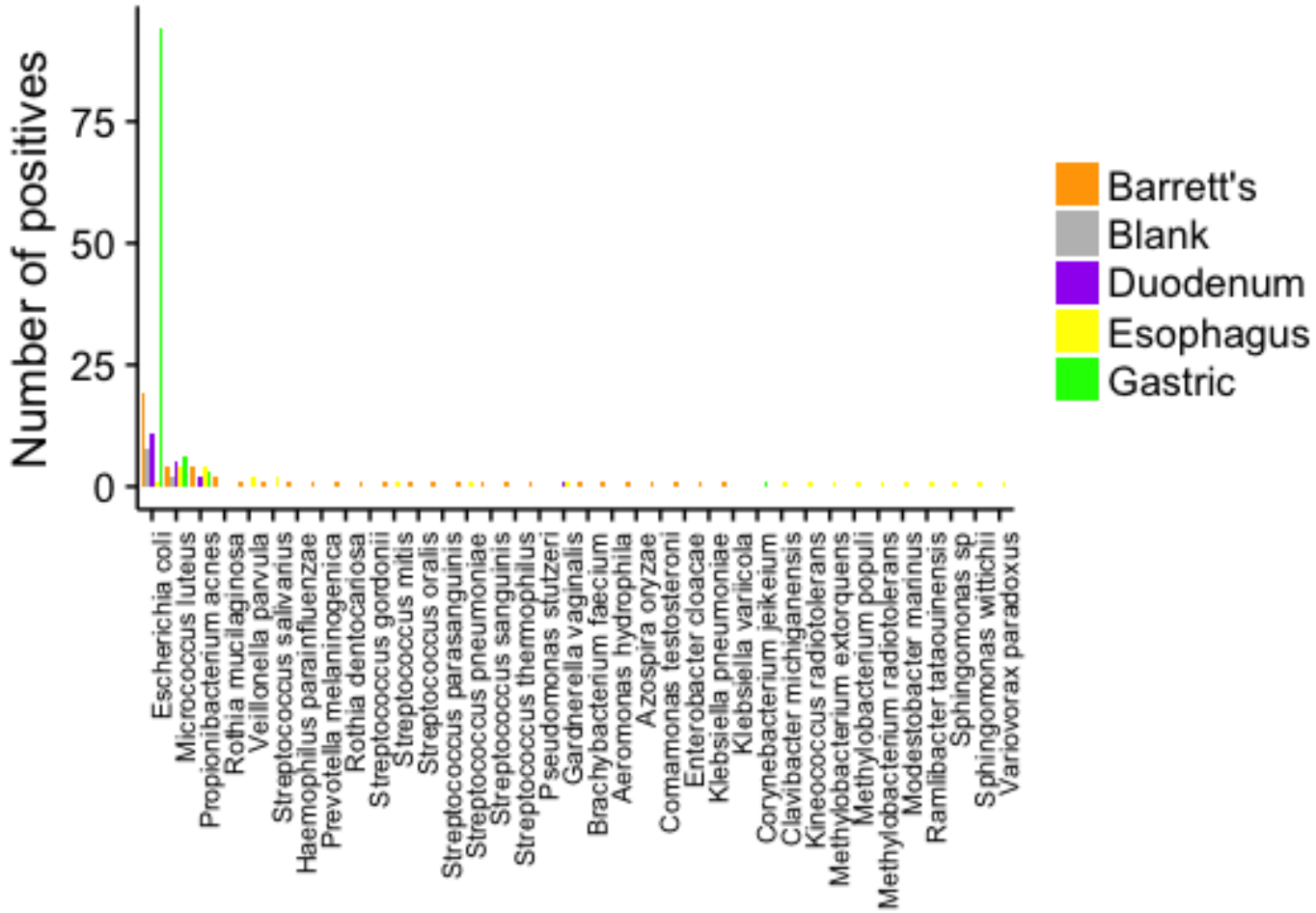


Figure A.1.53: Barplot depicts all pathogens and commensals detected by the poshTitan in single cell libraries generated from endoscopic biopsies taken from the proximal normal esophagus (yellow), stomach (green), duodenum (purple), and Barrett's lesion (orange). Brain whole tissue positive control (none detected) and blank water negative control (grey) samples were also analyzed. The number of samples in which the pathogen (x-axis) was detected is denoted by the height of the bar (y-axis) for tumor (dark grey) and normal (light grey) samples.

A.2 Supplementary Tables

A.2.1 Single cell Barrett's cohort descriptions

Patient	Age	Sex	Smoking (cigs/day)	Alcohol (units/week)	BMI	PPI	Prague
GEN02021	78	M	na	21	26.2	Yes	C3M5
GEN02023	77	M	na	4	25.7	Yes	C0M3
GEN02024	47	F	na	16	22.2	Yes	C0M2
GEN02025	90	F	na	6	23.7	Yes	C9M11
GEN02026	82	F	4	na	27.1	Yes	na
GEN02027	68	F	4	na	25.8	Yes	na

Table A.1: Relevant pathological and clinical data for the patients selected for single cell RNA-seq analysis. Patients were recruited and consented by Drs. Michael White and Richard Owen.

Batch	Patient	Tissue of origin	Sample No.	Clinical presentation
Batch A	GEN02021	Barrett's lesion	86	Barrett's oesophagus
Batch A	GEN02021	Prox. gastric	86	Barrett's oesophagus
Batch A	GEN02021	Prox. oesophagus	7	Barrett's oesophagus
Batch B	GEN02021	Barrett's lesion	171	Barrett's oesophagus
Batch B	GEN02021	Prox. gastric	46	Barrett's oesophagus
Batch B	Control	Brain total RNA (pos. cntrl)	2	na
Batch B	Control	Water blanks (neg. cntrl)	2	na
Batch C	GEN02021	Duodenum	90	Barrett's oesophagus
Batch C	GEN02021	Prox. gastric	90	Barrett's oesophagus
Batch C	GEN02023	Barrett's lesion	90	Barrett's oesophagus
Batch C	GEN02023	Prox. gastric	90	Barrett's oesophagus
Batch C	GEN02024	Barrett's lesion	90	Barrett's oesophagus
Batch C	GEN02024	Prox. oesophagus	90	Barrett's oesophagus
Batch C	GEN02025	Duodenum	90	Barrett's oesophagus
Batch C	GEN02025	Prox. oesophagus	90	Barrett's oesophagus
Batch C	Control	Brain total RNA (pos. cntrl)	24	na
Batch C	Control	Water blanks (neg. cntrl)	24	na
Batch D	GEN02021	Barrett's lesion	93	Barrett's oesophagus
Batch D	GEN02021	Prox. oesophagus	90	Barrett's oesophagus
Batch D	GEN02023	Duodenum	90	Barrett's oesophagus
Batch D	GEN02023	Prox. oesophagus	65	Barrett's oesophagus
Batch D	GEN02024	Duodenum	82	Barrett's oesophagus
Batch D	GEN02024	Prox. gastric	90	Barrett's oesophagus

Batch	Patient	Tissue of origin	Sample No.	Clinical presentation
Batch D	GEN02025	Barrett's lesion	90	Barrett's oesophagus
Batch D	GEN02025	Prox. gastric	90	Barrett's oesophagus
Batch D	Control	Brain total RNA (pos. cntrl)	24	na
Batch D	Control	Water blanks (neg. cntrl)	21	na
Batch E	GEN02021	Prox. oesophagus	92	Barrett's oesophagus
Batch E	GEN02023	Prox. oesophagus	92	Barrett's oesophagus
Batch E	GEN02024	Prox. gastric	93	Barrett's oesophagus
Batch E	GEN02025	Prox. gastric	93	Barrett's oesophagus
Batch E	Control	Brain total RNA (pos. cntrl)	11	na
Batch F	GEN02026	Norm. oesophagus	180	Gastric reflux only
Batch F	GEN02027	Norm. oesophagus	539	Gastric reflux only
Batch F	Control	Brain total RNA (pos. cntrl)	24	na
Batch F	Control	Water blanks (neg. cntrl)	24	na

Table A.2: Number of samples by batch, patient, and tissue. Endoscopic biopsies were collected and processed into lysed single cells by Drs. Michael White and Richard Owen.

A.2.2 BEARscc metrics and differential expression exclusion

Tissue	Cluster	Score	Stability	Promiscuity	Cluster size
Barrett's esophagus	1	0.887	0.942	0.056	75
Barrett's esophagus	2	0.647	0.980	0.332	35
Barrett's esophagus	3	0.103	0.582	0.479	10
Barrett's esophagus	4	0.921	0.953	0.032	201
Barrett's esophagus	5	0.692	0.978	0.286	50
Duodenum	1	0.765	1.00	0.235	6
Duodenum	2	0.731	0.773	0.042	40
Duodenum	3	0.922	0.941	0.019	94
Duodenum	4	0.787	0.801	0.014	47
Duodenum	5	0.852	0.905	0.054	20
Normal esophagus	1	0.997	0.998	0.000	762
Normal esophagus	2	0.830	0.996	0.166	89
Normal esophagus	3	0.830	1.000	0.170	88
Gastric	1	0.928	0.961	0.033	120
Gastric	2	0.903	0.957	0.055	79
Gastric	3	0.958	0.970	0.013	22

Table A.3: BEARscc metrics for the chosen BEARscc clusters of the four endsocopic biopsy tissues: Barrett's esophagus, duodenum, normal esophagus, and gastric.

Cluster	No. cells included for DE	No. cells excluded for DE	Score threshold	Cluster size
1	754	8	0.95	762
2	88	1	0.82	89
3	88	0	0.82	88

Table A.4: Table displays the number of cells excluded from or included in differential expression analysis for each esophageal BEARscc cluster. The BEARscc score threshold applied to determine inclusion in differential expression analysis is also indicated in the table in addition to cluster size.

Cluster	No. cells included for DE	No. cells excluded for DE	Score threshold	Cluster size
1	109	11	0.90	120
2	69	10	0.90	79
3	20	2	0.90	22

Table A.5: Table displays the number of cells excluded from or included in differential expression analysis for each gastric BEARscc cluster. The BEARscc score threshold applied to determine inclusion in differential expression analysis is also indicated in the table in addition to cluster size.

Cluster	No. cells included for DE	No. cells excluded for DE	Score threshold	Cluster size
1	6	0	na	6
2	24	16	0.82	40
3	78	16	0.94	94
4	38	9	0.88	47
5	11	9	0.9	20

Table A.6: Table displays the number of cells excluded from or included in differential expression analysis for each duodenal BEARscc cluster. The BEARscc score threshold applied to determine inclusion in differential expression analysis is also indicated in the table in addition to cluster size.

Cluster	No. cells included for DE	No. cells excluded for DE	Score threshold	Cluster size
1	66	9	0.92	75
2	31	4	0.65	35
3	3	7	0.10	10
4	175	26	0.94	201
5	40	10	0.69	50

Table A.7: Table displays the number of cells excluded from or included in differential expression analysis for each Barrett's tissue BEARscc cluster. The BEARscc score threshold applied to determine inclusion in differential expression analysis is also indicated in the table in addition to cluster size.

Cluster	No. cells included for DE	No. cells excluded for DE	Score threshold	Cluster size
1	421	95	0.81	516
2	254	38 4	0.89	292
3	177	11	0.98	188
4	216	79	0.88	295
5	152	14	0.97	166
6a	58	4	0.32	62
6b	86	3	0.45	89
6c	3	0	na	3
6d	83	0	0.55	83
6e	1	0	na	1
6f	1	0	na	1
7	102	23	0.85	125

Table A.8: Table displays the number of cells excluded from or included in differential expression analysis for each single cell BEARscc cluster. The BEARscc score threshold applied to determine inclusion in differential expression analysis is also indicated in the table in addition to cluster size.

A.2.3 Differentially expressed genes across the esophageal clusters

Table A.9 Esophageal differentially expressed genes.

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000170477	KRT4	6.71	3.99e-67	0.000
1	ENSG00000229807	XIST	6.63	1.17e-84	0.000
1	ENSG00000178934	LGALS7B	5.37	0.000	0.000
1	ENSG00000256812	CAPNS2	4.73	4.36e-67	0.000
1	ENSG00000196344	ADH7	4.3	1.8e-89	0.000
1	ENSG00000121552	CSTA	3.96	1.14e-138	0.000
1	ENSG00000232346	ENSG00000232346	3.65	9.61e-117	0.000
1	ENSG00000209082	MT-TL1	3.3	3.33e-56	0.000
1	ENSG00000249119	MTND6P4	3.26	1.86e-53	0.000
1	ENSG00000219928	ENSG00000219928	3.15	1.23e-66	0.000
1	ENSG00000181524	RPL24P4	3.07	4e-90	0.000
1	ENSG00000237991	RPL35P1	3.05	1.14e-63	0.000
1	ENSG00000218175	ENSG00000218175	3.01	7.85e-177	0.000
1	ENSG00000163331	DAPL1	2.65	3.32e-64	0.000
1	ENSG00000231767	ENSG00000231767	2.64	1.12e-114	0.000
1	ENSG00000228205	ENSG00000228205	2.63	1.08e-101	0.000
1	ENSG00000235174	RPL39P3	2.35	1.32e-66	0.000
1	ENSG00000234964	FABP5P7	2.24	8.02e-123	0.000
1	ENSG00000146677	ENSG00000146677	2.16	1.36e-104	0.000
1	ENSG00000232472	EEF1B2P3	2.08	1.86e-86	0.000
1	ENSG00000227615	ENSG00000227615	2.05	1.36e-67	0.000
1	ENSG00000224631	ENSG00000224631	2.02	3.59e-106	0.000
1	ENSG00000220472	ENSG00000220472	1.98	1.81e-70	0.000
1	ENSG00000178429	RPS3AP5	1.97	6.25e-53	0.000
1	ENSG00000228929	RPS13P2	1.97	6.73e-59	0.000
1	ENSG00000214485	RPL7P1	1.93	1.23e-113	0.000
1	ENSG00000213598	ENSG00000213598	1.89	6.18e-77	0.000
1	ENSG00000226221	RPL26P19	1.87	3.58e-55	0.000
1	ENSG00000218426	ENSG00000218426	1.8	3.82e-108	0.000
1	ENSG00000230629	RPS23P8	1.79	1.4e-57	0.000
1	ENSG00000139239	RPL14P1	1.62	1.46e-65	0.000
1	ENSG00000187193	MT1X	1.61	4.82e-59	0.000
1	ENSG00000226084	ENSG00000226084	1.59	2.88e-60	0.000
1	ENSG00000229638	RPL4P4	1.58	6.63e-80	0.000
1	ENSG00000226360	RPL10AP6	1.57	1.71e-61	0.000
1	ENSG00000136810	TXN	1.4	7.53e-76	0.000
1	ENSG00000213553	RPLP0P6	1.4	6.12e-98	0.000
1	ENSG00000233762	ENSG00000233762	1.32	2.38e-54	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000214389	RPS3AP26	1.24	2.07e-52	0.000
1	ENSG00000236552	RPL13AP5	1.03	2.08e-62	0.000
1	ENSG00000244122	UGT1A7	4.78	1.43e-48	0.002
1	ENSG00000236801	RPL24P8	2.3	1.11e-47	0.002
1	ENSG00000235043	TECRP1	2.05	1.42e-46	0.002
1	ENSG00000243199	ENSG00000243199	1.51	3.78e-51	0.002
1	ENSG00000212802	RPL15P3	1.15	1.91e-49	0.002
1	ENSG00000224550	ENSG00000224550	2.95	2.55e-46	0.003
1	ENSG00000226608	FTLP3	2.33	3.77e-45	0.004
1	ENSG00000229117	RPL41	0.675	8.12e-46	0.004
1	ENSG00000164825	DEFB1	3.42	3e-44	0.007
1	ENSG00000225067	RPL23AP2	2.58	1.63e-44	0.007
1	ENSG00000185641	ENSG00000185641	1.39	6.21e-44	0.007
1	ENSG00000235084	CHCHD2P6	2.54	1.42e-43	0.008
1	ENSG00000230897	RPS18P12	2.26	1.89e-43	0.008
1	ENSG00000237214	ENSG00000237214	1.96	6.85e-43	0.011
1	ENSG00000234036	TXNP6	2.93	1.29e-42	0.014
1	ENSG00000242299	ENSG00000242299	1.05	3.6e-42	0.019
1	ENSG00000244313	ENSG00000244313	1.1	7.73e-42	0.021
1	ENSG00000227097	RPS28P7	0.975	1.75e-40	0.045
1	ENSG00000197956	S100A6	-1.56	3.35e-79	0.000
1	ENSG00000206503	HLA-A	-1.57	1.69e-52	0.000
1	ENSG00000234745	HLA-B	-1.64	1.01e-54	0.000
1	ENSG00000263563	UBBP4	-2.12	6.49e-61	0.000
1	ENSG00000204525	HLA-C	-2.29	1.69e-86	0.000
1	ENSG00000204632	HLA-G	-2.37	1.62e-56	0.000
1	ENSG00000206341	HLA-H	-2.44	7.11e-74	0.000
1	ENSG00000111057	KRT18	-2.82	2.08e-58	0.000
1	ENSG00000224520	KRT8P45	-2.95	2.13e-65	0.000
1	ENSG00000204622	HLA-J	-2.98	8.02e-104	0.000
1	ENSG00000271043	MTRNR2L2	-2.99	2.62e-74	0.000
1	ENSG00000169218	RSPO1	-3.14	3.36e-72	0.000
1	ENSG00000254285	KRT8P3	-3.33	1.58e-62	0.000
1	ENSG00000272763	ENSG00000272763	-3.43	2.58e-52	0.000
1	ENSG00000162779	AXDND1	-3.53	6.53e-55	0.000
1	ENSG00000146038	DCDC2	-3.61	4.3e-53	0.000
1	ENSG00000109265	KIAA1211	-3.63	9.98e-55	0.000
1	ENSG00000138100	TRIM54	-3.71	5.52e-58	0.000
1	ENSG00000182333	LIPF	-3.75	3.85e-60	0.000
1	ENSG00000099617	EFNA2	-3.82	5.72e-59	0.000
1	ENSG00000005073	HOXA11	-3.82	3.03e-54	0.000
1	ENSG00000205922	ONECUT3	-3.82	3.94e-59	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000157399	ARSE	-3.85	2.96e-52	0.000
1	ENSG00000117122	MFAP2	-3.92	1e-52	0.000
1	ENSG00000114248	LRRC31	-3.94	6.15e-64	0.000
1	ENSG00000224081	LINC01057	-3.94	7.39e-56	0.000
1	ENSG00000134339	SAA2	-3.95	2.6e-73	0.000
1	ENSG00000168903	BTNL3	-3.97	1.1e-52	0.000
1	ENSG00000266010	GATA6-AS1	-4.03	8.35e-55	0.000
1	ENSG00000134871	COL4A2	-4.07	4.06e-59	0.000
1	ENSG00000170421	KRT8	-4.07	1.28e-73	0.000
1	ENSG00000122679	RAMP3	-4.11	8.3e-57	0.000
1	ENSG00000226025	ENSG00000226025	-4.14	3.49e-65	0.000
1	ENSG00000130035	GALNT8	-4.15	2.33e-52	0.000
1	ENSG00000125384	PTGER2	-4.17	6.33e-60	0.000
1	ENSG00000019582	CD74	-4.22	9.29e-57	0.000
1	ENSG00000180251	SLC9A4	-4.26	1.52e-56	0.000
1	ENSG00000240990	HOXA11-AS	-4.29	5.85e-54	0.000
1	ENSG00000163501	IHH	-4.29	2.08e-64	0.000
1	ENSG00000198535	C2CD4A	-4.3	6.45e-69	0.000
1	ENSG00000168062	BATF2	-4.34	1.55e-62	0.000
1	ENSG00000213886	UBD	-4.35	1.87e-79	0.000
1	ENSG00000240668	KRT8P36	-4.36	2.86e-151	0.000
1	ENSG00000159184	HOXB13	-4.37	1.14e-63	0.000
1	ENSG00000156127	BATF	-4.37	4.17e-65	0.000
1	ENSG00000140968	IRF8	-4.39	2.23e-62	0.000
1	ENSG00000112818	MEP1A	-4.41	9.36e-57	0.000
1	ENSG00000145384	FABP2	-4.42	5.27e-58	0.000
1	ENSG00000108244	KRT23	-4.43	1.37e-59	0.000
1	ENSG00000205076	LGALS7	-4.44	1.1e-136	0.000
1	ENSG00000111110	PPM1H	-4.49	3.91e-69	0.000
1	ENSG00000106069	CHN2	-4.51	3.77e-55	0.000
1	ENSG00000015520	NPC1L1	-4.55	4.44e-56	0.000
1	ENSG00000204610	TRIM15	-4.55	1.11e-69	0.000
1	ENSG00000133019	CHRM3	-4.57	1.25e-60	0.000
1	ENSG00000118017	A4GNT	-4.59	1.22e-55	0.000
1	ENSG00000118849	RARRES1	-4.6	1.9e-56	0.000
1	ENSG00000139292	LGR5	-4.61	7.7e-54	0.000
1	ENSG00000123843	C4BPB	-4.63	2.31e-66	0.000
1	ENSG00000163817	SLC6A20	-4.69	3.56e-53	0.000
1	ENSG00000178882	FAM101A	-4.69	1.27e-65	0.000
1	ENSG00000099834	CDHR5	-4.69	2.4e-68	0.000
1	ENSG00000162692	VCAM1	-4.7	7.32e-64	0.000
1	ENSG00000165507	C10orf10	-4.72	1.69e-52	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000160188	RSPH1	-4.72	2.04e-58	0.000
1	ENSG00000121380	BCL2L14	-4.74	9.59e-62	0.000
1	ENSG00000153551	CMTM7	-4.75	1.25e-62	0.000
1	ENSG00000243766	HOTTIP	-4.75	9.05e-53	0.000
1	ENSG00000136052	SLC41A2	-4.76	2.5e-63	0.000
1	ENSG00000188993	LRRC66	-4.76	7.33e-60	0.000
1	ENSG00000120075	HOXB5	-4.79	6.42e-74	0.000
1	ENSG00000131910	NR0B2	-4.79	1.36e-61	0.000
1	ENSG00000113303	BTNL8	-4.79	3.69e-70	0.000
1	ENSG00000028137	TNFRSF1B	-4.8	3.44e-65	0.000
1	ENSG00000067646	ZFY	-4.8	1.24e-63	0.000
1	ENSG00000055163	CYFIP2	-4.81	4.83e-65	0.000
1	ENSG00000115616	SLC9A2	-4.84	1.97e-64	0.000
1	ENSG00000178401	DNAJC22	-4.84	2.09e-65	0.000
1	ENSG00000179674	ARL14	-4.85	2.76e-65	0.000
1	ENSG00000204019	CT83	-4.89	1.07e-82	0.000
1	ENSG00000232629	HLA-DQB2	-4.9	8.69e-130	0.000
1	ENSG00000166866	MYO1A	-4.91	5.91e-63	0.000
1	ENSG00000009765	IYD	-4.93	1.67e-68	0.000
1	ENSG00000100170	SLC5A1	-4.95	5.13e-61	0.000
1	ENSG00000100302	RASD2	-4.97	1.69e-59	0.000
1	ENSG00000021300	PLEKHB1	-4.97	2.94e-57	0.000
1	ENSG00000008517	IL32	-4.99	8.59e-56	0.000
1	ENSG00000225630	MTND2P28	-5	7.42e-191	0.000
1	ENSG00000261857	MIA	-5.01	1.31e-90	0.000
1	ENSG00000166920	C15orf48	-5.01	1.17e-53	0.000
1	ENSG00000204876	ENSG00000204876	-5.01	2.63e-72	0.000
1	ENSG00000233864	TTTY15	-5.01	2.87e-113	0.000
1	ENSG00000107242	PIP5K1B	-5.02	3.46e-81	0.000
1	ENSG00000183018	SPNS2	-5.03	4.34e-84	0.000
1	ENSG00000163515	RETNLB	-5.05	8.67e-60	0.000
1	ENSG00000162174	ASRGL1	-5.06	1.15e-74	0.000
1	ENSG00000198846	TOX	-5.07	3.46e-74	0.000
1	ENSG00000136574	GATA4	-5.09	6.23e-85	0.000
1	ENSG00000204616	TRIM31	-5.09	2.69e-59	0.000
1	ENSG00000166869	CHP2	-5.12	1.15e-53	0.000
1	ENSG00000106031	HOXA13	-5.13	1.43e-81	0.000
1	ENSG00000186854	TRABD2A	-5.14	1.43e-56	0.000
1	ENSG00000223865	HLA-DPB1	-5.14	6.41e-53	0.000
1	ENSG00000196436	NPIPB15	-5.14	1.7e-83	0.000
1	ENSG00000170608	FOXA3	-5.17	8.15e-89	0.000
1	ENSG00000164749	HNF4G	-5.17	4.17e-77	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000148702	HABP2	-5.17	2.53e-64	0.000
1	ENSG00000229391	HLA-DRB6	-5.18	1.29e-203	0.000
1	ENSG00000138722	MMRN1	-5.2	7.18e-71	0.000
1	ENSG00000183508	FAM46C	-5.21	2.47e-54	0.000
1	ENSG00000123358	NR4A1	-5.21	2.1e-61	0.000
1	ENSG00000139540	SLC39A5	-5.21	3.03e-71	0.000
1	ENSG00000163072	NOSTRIN	-5.22	1.07e-80	0.000
1	ENSG00000009950	MLXIPL	-5.24	1.96e-52	0.000
1	ENSG00000108511	HOXB6	-5.26	5.97e-94	0.000
1	ENSG00000139515	PDX1	-5.27	7.71e-80	0.000
1	ENSG00000112699	GMDS	-5.28	8.79e-60	0.000
1	ENSG00000104537	ANXA13	-5.29	1.81e-72	0.000
1	ENSG00000163586	FABP1	-5.29	6.14e-56	0.000
1	ENSG00000114374	USP9Y	-5.29	2.39e-105	0.000
1	ENSG00000067048	DDX3Y	-5.31	2.53e-52	0.000
1	ENSG00000187608	ISG15	-5.32	2.48e-65	0.000
1	ENSG00000196730	DAPK1	-5.32	3.18e-78	0.000
1	ENSG00000170439	METTL7B	-5.32	2.41e-66	0.000
1	ENSG00000101577	LPIN2	-5.34	1.13e-76	0.000
1	ENSG00000166510	CCDC68	-5.34	5.84e-84	0.000
1	ENSG00000174358	SLC6A19	-5.36	2.49e-69	0.000
1	ENSG00000197635	DPP4	-5.36	2.04e-88	0.000
1	ENSG00000163739	CXCL1	-5.38	4.67e-74	0.000
1	ENSG00000125798	FOXA2	-5.4	2.24e-81	0.000
1	ENSG00000165556	CDX2	-5.4	6.51e-82	0.000
1	ENSG00000161798	AQP5	-5.42	3.66e-82	0.000
1	ENSG00000135773	CAPN9	-5.42	1.16e-79	0.000
1	ENSG00000184012	TMPRSS2	-5.45	6.04e-54	0.000
1	ENSG00000169562	GJB1	-5.45	1.22e-101	0.000
1	ENSG00000142484	TM4SF5	-5.46	6.01e-93	0.000
1	ENSG00000154451	GBP5	-5.48	1.14e-63	0.000
1	ENSG00000150594	ADRA2A	-5.52	1.42e-85	0.000
1	ENSG00000162078	ZG16B	-5.53	6.94e-85	0.000
1	ENSG00000086696	HSD17B2	-5.53	4.14e-68	0.000
1	ENSG00000169248	CXCL11	-5.56	4.76e-101	0.000
1	ENSG00000196735	HLA-DQA1	-5.58	3.33e-58	0.000
1	ENSG00000198488	B3GNT6	-5.6	2.12e-72	0.000
1	ENSG00000132437	DDC	-5.61	3.13e-85	0.000
1	ENSG00000198502	HLA-DRB5	-5.66	7.94e-58	0.000
1	ENSG00000053918	KCNQ1	-5.66	1.37e-78	0.000
1	ENSG00000070019	GUCY2C	-5.68	8.85e-72	0.000
1	ENSG00000198734	F5	-5.68	5.81e-82	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000012171	SEMA3B	-5.69	6.03e-77	0.000
1	ENSG00000137959	IFI44L	-5.69	3.03e-60	0.000
1	ENSG00000169347	GP2	-5.73	3.08e-75	0.000
1	ENSG00000002933	TMEM176A	-5.73	8.5e-84	0.000
1	ENSG00000132911	NMUR2	-5.74	7.5e-67	0.000
1	ENSG00000117707	PROX1	-5.74	2.1e-77	0.000
1	ENSG00000160183	TMPRSS3	-5.75	3.14e-90	0.000
1	ENSG00000123191	ATP7B	-5.75	6.15e-72	0.000
1	ENSG00000126709	IFI6	-5.76	2.68e-82	0.000
1	ENSG00000250337	LINC01021	-5.77	3.21e-101	0.000
1	ENSG00000257743	MGAM2	-5.77	7.18e-87	0.000
1	ENSG00000255545	ENSG00000255545	-5.79	3.98e-68	0.000
1	ENSG00000183778	B3GALT5	-5.8	3.46e-73	0.000
1	ENSG00000113722	CDX1	-5.81	2.64e-79	0.000
1	ENSG00000096395	MLN	-5.81	5.12e-72	0.000
1	ENSG00000074276	CDHR2	-5.83	4.88e-84	0.000
1	ENSG00000113924	HGD	-5.87	2.54e-102	0.000
1	ENSG00000148584	A1CF	-5.87	8.3e-92	0.000
1	ENSG00000178828	RNF186	-5.87	3.74e-75	0.000
1	ENSG00000168631	DPCR1	-5.9	8.98e-65	0.000
1	ENSG00000144820	ADGRG7	-5.92	5.25e-135	0.000
1	ENSG00000019102	VSIG2	-5.93	1.05e-60	0.000
1	ENSG00000183878	UTY	-5.94	3.53e-149	0.000
1	ENSG00000104783	KCNN4	-5.95	2.53e-71	0.000
1	ENSG00000198692	EIF1AY	-5.95	4.07e-212	0.000
1	ENSG00000101842	VSIG1	-5.97	3.72e-62	0.000
1	ENSG00000012817	KDM5D	-5.98	7.2e-108	0.000
1	ENSG00000141448	GATA6	-5.98	4.79e-135	0.000
1	ENSG00000163898	LIPH	-5.99	3.16e-85	0.000
1	ENSG00000099725	PRKY	-6.02	4.04e-114	0.000
1	ENSG00000120875	DUSP4	-6.03	1.05e-102	0.000
1	ENSG00000180745	CLRN3	-6.03	4.16e-113	0.000
1	ENSG00000135480	KRT7	-6.04	1.55e-56	0.000
1	ENSG00000148735	PLEKHS1	-6.05	1.93e-88	0.000
1	ENSG00000106331	PAX4	-6.06	7.07e-64	0.000
1	ENSG00000125730	C3	-6.07	1.01e-79	0.000
1	ENSG00000107159	CA9	-6.07	1.83e-82	0.000
1	ENSG00000131771	PPP1R1B	-6.1	6.14e-57	0.000
1	ENSG00000088992	TESC	-6.1	3.28e-85	0.000
1	ENSG00000184709	LRRC26	-6.11	6.22e-76	0.000
1	ENSG00000166922	SCG5	-6.12	1.86e-71	0.000
1	ENSG00000169994	MYO7B	-6.13	1.04e-83	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000101194	SLC17A9	-6.14	5.8e-88	0.000
1	ENSG00000198643	FAM3D	-6.16	5.81e-58	0.000
1	ENSG00000169213	RAB3B	-6.17	8.89e-78	0.000
1	ENSG00000160862	AZGP1	-6.19	2.71e-84	0.000
1	ENSG00000156966	B3GNT7	-6.21	1.18e-97	0.000
1	ENSG00000042980	ADAM28	-6.22	2.61e-102	0.000
1	ENSG00000096088	PGC	-6.23	4.15e-102	0.000
1	ENSG00000101333	PLCB4	-6.24	2.47e-104	0.000
1	ENSG00000158764	ITLN2	-6.24	9.55e-66	0.000
1	ENSG00000119547	ONECUT2	-6.24	8.74e-80	0.000
1	ENSG00000267795	SMIM22	-6.25	2.87e-101	0.000
1	ENSG00000169507	SLC38A11	-6.27	3.49e-88	0.000
1	ENSG00000236699	ARHGEF38	-6.27	2.47e-120	0.000
1	ENSG00000106565	TMEM176B	-6.27	7.01e-90	0.000
1	ENSG00000135052	GOLM1	-6.28	3.39e-66	0.000
1	ENSG00000089472	HEPH	-6.28	8.28e-93	0.000
1	ENSG00000103460	TOX3	-6.28	5.45e-122	0.000
1	ENSG00000173597	SULT1B1	-6.28	5.54e-90	0.000
1	ENSG00000176907	C8orf4	-6.29	5.77e-107	0.000
1	ENSG00000162878	PKDCC	-6.3	1.1e-106	0.000
1	ENSG00000162545	CAMK2N1	-6.3	2.65e-115	0.000
1	ENSG00000081041	CXCL2	-6.36	1.13e-77	0.000
1	ENSG00000137501	SYTL2	-6.36	1.02e-69	0.000
1	ENSG00000101076	HNF4A	-6.37	8.69e-128	0.000
1	ENSG00000112379	ARFGEF3	-6.38	3.12e-76	0.000
1	ENSG00000179913	B3GNT3	-6.39	7.93e-103	0.000
1	ENSG00000157005	SST	-6.41	3.06e-74	0.000
1	ENSG00000203697	CAPN8	-6.42	3.5e-71	0.000
1	ENSG00000013588	GPRC5A	-6.42	9.43e-57	0.000
1	ENSG00000231389	HLA-DPA1	-6.43	4.1e-63	0.000
1	ENSG00000143167	GPA33	-6.43	1.26e-88	0.000
1	ENSG00000074370	ATP2A3	-6.45	1.53e-82	0.000
1	ENSG00000122176	FMOD	-6.47	3.25e-87	0.000
1	ENSG00000122859	NEUROG3	-6.51	6.52e-74	0.000
1	ENSG00000102854	MSLN	-6.54	3.7e-107	0.000
1	ENSG00000198758	EPS8L3	-6.55	7.11e-95	0.000
1	ENSG00000010438	PRSS3	-6.55	4.25e-82	0.000
1	ENSG00000115386	REG1A	-6.57	3.11e-77	0.000
1	ENSG00000171431	KRT20	-6.6	1.07e-70	0.000
1	ENSG00000165140	FBP1	-6.62	6.84e-136	0.000
1	ENSG00000233041	PHGR1	-6.64	7.18e-130	0.000
1	ENSG00000164822	DEFA6	-6.64	5.11e-70	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000145113	MUC4	-6.65	4.5e-77	0.000
1	ENSG00000163734	CXCL3	-6.65	1.28e-105	0.000
1	ENSG00000124664	SPDEF	-6.66	2.77e-88	0.000
1	ENSG00000140297	GCNT3	-6.75	1.14e-75	0.000
1	ENSG00000166959	MS4A8	-6.75	1.34e-83	0.000
1	ENSG00000214814	FER1L6	-6.77	2.27e-110	0.000
1	ENSG00000204252	HLA-DOA	-6.77	3.31e-124	0.000
1	ENSG00000095932	SMIM24	-6.78	6.29e-145	0.000
1	ENSG00000117525	F3	-6.78	3.26e-94	0.000
1	ENSG00000166825	ANPEP	-6.78	7.7e-104	0.000
1	ENSG00000164816	DEFA5	-6.8	9.33e-74	0.000
1	ENSG00000127831	VIL1	-6.88	2.51e-101	0.000
1	ENSG00000173432	SAA1	-6.88	1.77e-111	0.000
1	ENSG00000099812	MISP	-6.9	3.12e-135	0.000
1	ENSG00000006611	USH1C	-6.94	9.06e-147	0.000
1	ENSG00000002726	AOC1	-6.98	6.85e-111	0.000
1	ENSG00000160181	TFF2	-6.99	2.96e-100	0.000
1	ENSG00000137860	SLC28A2	-7.03	7.47e-110	0.000
1	ENSG00000169894	MUC3A	-7.05	4.19e-112	0.000
1	ENSG00000129538	RNASE1	-7.06	9.19e-104	0.000
1	ENSG00000243709	LEFTY1	-7.1	2.54e-76	0.000
1	ENSG00000129824	RPS4Y1	-7.15	1.02e-214	0.000
1	ENSG00000130513	GDF15	-7.15	2.28e-97	0.000
1	ENSG00000145703	IQGAP2	-7.15	1.42e-202	0.000
1	ENSG00000165215	CLDN3	-7.17	3.19e-123	0.000
1	ENSG00000079112	CDH17	-7.21	3.17e-137	0.000
1	ENSG00000259974	LINC00261	-7.21	1.49e-64	0.000
1	ENSG00000179914	ITLN1	-7.22	4.89e-68	0.000
1	ENSG00000131002	TXLNGY	-7.23	2.29e-293	0.000
1	ENSG00000188175	HEPACAM2	-7.24	4.62e-85	0.000
1	ENSG00000131203	IDO1	-7.25	1.28e-148	0.000
1	ENSG00000175130	MARCKSL1	-7.36	1.9e-183	0.000
1	ENSG00000007062	PROM1	-7.36	2.75e-158	0.000
1	ENSG00000064787	BCAS1	-7.41	5.79e-121	0.000
1	ENSG00000169876	MUC17	-7.43	1.35e-89	0.000
1	ENSG00000204385	SLC44A4	-7.5	3.42e-116	0.000
1	ENSG00000127324	TSPAN8	-7.55	9.16e-89	0.000
1	ENSG00000103534	TMC5	-7.57	5.21e-139	0.000
1	ENSG00000115648	MLPH	-7.66	1.06e-136	0.000
1	ENSG00000198203	SULT1C2	-7.83	7.57e-129	0.000
1	ENSG00000066405	CLDN18	-7.83	2.18e-133	0.000
1	ENSG00000185499	MUC1	-7.85	1.14e-132	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000134398	ERN2	-7.89	9.42e-155	0.000
1	ENSG00000138755	CXCL9	-7.89	8.24e-172	0.000
1	ENSG00000157613	CREB3L1	-7.9	3.27e-180	0.000
1	ENSG00000173702	MUC13	-7.92	4.83e-164	0.000
1	ENSG00000169245	CXCL10	-7.96	6.99e-146	0.000
1	ENSG00000173467	AGR3	-8.12	2.45e-174	0.000
1	ENSG00000101443	WFDC2	-8.41	7.88e-206	0.000
1	ENSG00000105388	CEACAM5	-8.46	1.67e-91	0.000
1	ENSG00000109511	ANXA10	-8.51	4.24e-165	0.000
1	ENSG00000116299	KIAA1324	-8.62	1.3e-175	0.000
1	ENSG00000122711	SPINK4	-8.68	9.41e-122	0.000
1	ENSG00000196188	CTSE	-8.7	3.54e-168	0.000
1	ENSG00000117983	MUC5B	-8.73	8.66e-134	0.000
1	ENSG00000163993	S100P	-8.8	7.91e-173	0.000
1	ENSG00000164266	SPINK1	-8.91	1.65e-210	0.000
1	ENSG00000162896	PIGR	-8.92	3.24e-129	0.000
1	ENSG00000198788	MUC2	-8.93	1.1e-92	0.000
1	ENSG00000171747	LGALS4	-8.95	1.23e-138	0.000
1	ENSG00000187908	DMBT1	-8.96	1.28e-147	0.000
1	ENSG00000106541	AGR2	-9.01	2.27e-113	0.000
1	ENSG00000160180	TFF3	-9.36	6.12e-140	0.000
1	ENSG00000090382	LYZ	-9.86	1.51e-164	0.000
1	ENSG00000197249	SERPINA1	-9.89	9.39e-201	0.000
1	ENSG00000102837	OLFM4	-10.1	1.78e-159	0.000
1	ENSG00000160182	TFF1	-10.3	5.05e-211	0.000
1	ENSG00000134193	REG4	-10.6	1.11e-173	0.000
1	ENSG00000184956	MUC6	-11.1	1.71e-166	0.000
1	ENSG00000215182	MUC5AC	-11.8	7.09e-277	0.000
1	ENSG00000109193	SULT1E1	-3.62	5.24e-52	0.001
1	ENSG00000176728	TTTY14	-3.72	6.8e-52	0.001
1	ENSG00000137251	TINAG	-3.77	7.95e-52	0.001
1	ENSG00000129757	CDKN1C	-3.84	5.65e-52	0.001
1	ENSG00000016082	ISL1	-4.06	6.42e-52	0.001
1	ENSG00000248596	ENSG00000248596	-4.39	6.3e-52	0.001
1	ENSG00000188833	ENTPD8	-4.63	8.32e-52	0.001
1	ENSG00000088340	FER1L4	-5	7.18e-52	0.001
1	ENSG00000229031	ENSG00000229031	-2.8	1.03e-50	0.002
1	ENSG00000154620	TMSB4Y	-3.27	2.16e-49	0.002
1	ENSG00000149948	HMGA2	-3.43	3.02e-49	0.002
1	ENSG00000117152	RGS4	-3.51	6.45e-47	0.002
1	ENSG00000116574	RHOU	-3.52	4.83e-49	0.002
1	ENSG00000111701	APOBEC1	-3.54	5.07e-50	0.002

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000251350	ENSG00000251350	-3.57	4.58e-48	0.002
1	ENSG00000130957	FBP2	-3.65	2.55e-47	0.002
1	ENSG00000213996	TM6SF2	-3.67	8.26e-47	0.002
1	ENSG00000144852	NR1I2	-3.67	2.12e-48	0.002
1	ENSG00000138207	RBP4	-3.7	1.17e-46	0.002
1	ENSG00000105251	SHD	-3.71	3.57e-50	0.002
1	ENSG00000150750	C11orf53	-3.76	1.51e-49	0.002
1	ENSG00000114455	HHLA2	-3.78	2.56e-47	0.002
1	ENSG00000166143	PPP1R14D	-3.86	1.23e-48	0.002
1	ENSG00000090402	SI	-3.89	7.61e-51	0.002
1	ENSG00000173237	C11orf86	-3.94	2.08e-49	0.002
1	ENSG00000198286	CARD11	-3.95	2.39e-50	0.002
1	ENSG00000204381	LAYN	-3.95	3.36e-51	0.002
1	ENSG00000165626	BEND7	-3.96	1.88e-48	0.002
1	ENSG00000146039	SLC17A4	-3.97	5.04e-48	0.002
1	ENSG00000123143	PKN1	-3.97	1.02e-47	0.002
1	ENSG00000119888	EPCAM	-4.05	5.6e-47	0.002
1	ENSG00000152503	TRIM36	-4.06	1.18e-46	0.002
1	ENSG00000184347	SLIT3	-4.17	4.38e-50	0.002
1	ENSG00000169903	TM4SF4	-4.18	1.09e-48	0.002
1	ENSG00000185338	SOCS1	-4.18	3.78e-51	0.002
1	ENSG00000197410	DCHS2	-4.24	2.3e-49	0.002
1	ENSG00000183960	KCNH8	-4.27	4.72e-47	0.002
1	ENSG00000101670	LIPG	-4.34	4.58e-50	0.002
1	ENSG00000128917	DLL4	-4.36	8.21e-48	0.002
1	ENSG00000176532	PRR15	-4.39	2.34e-47	0.002
1	ENSG00000175311	ANKS4B	-4.4	1.15e-47	0.002
1	ENSG00000248771	LINC01207	-4.53	2.06e-46	0.002
1	ENSG00000187678	SPRY4	-4.59	4.08e-50	0.002
1	ENSG00000106772	PRUNE2	-4.6	1.68e-50	0.002
1	ENSG00000135114	OASL	-4.66	6.34e-47	0.002
1	ENSG00000142185	TRPM2	-4.89	2.45e-48	0.002
1	ENSG00000169116	PARM1	-4.94	7.07e-49	0.002
1	ENSG00000204128	C2orf72	-5	2.24e-50	0.002
1	ENSG00000165376	CLDN2	-5.03	1.96e-48	0.002
1	ENSG00000001626	CFTR	-5.04	1.25e-46	0.002
1	ENSG00000129946	SHC2	-5.04	1.36e-50	0.002
1	ENSG00000250069	ENSG00000250069	-5.06	2.12e-49	0.002
1	ENSG00000100604	CHGA	-5.13	4.63e-51	0.002
1	ENSG00000108551	RASD1	-5.21	3.07e-47	0.002
1	ENSG00000242574	HLA-DMB	-5.23	1.68e-50	0.002
1	ENSG00000172482	AGXT	-5.25	4.12e-49	0.002

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000143341	HMCN1	-5.31	4.28e-50	0.002
1	ENSG00000177685	CRACR2B	-5.36	7.97e-51	0.002
1	ENSG00000267541	MTCO2P2	-2.89	6.95e-46	0.003
1	ENSG00000124766	SOX4	-3.1	4.28e-46	0.003
1	ENSG00000165246	NLGN4Y	-3.16	2.66e-46	0.003
1	ENSG00000180818	HOXC10	-3.94	7.07e-46	0.003
1	ENSG00000160991	ORAI2	-4.06	3.71e-46	0.003
1	ENSG00000126218	F10	-4.1	4.7e-46	0.003
1	ENSG00000105707	HPN	-4.48	4.07e-46	0.003
1	ENSG00000172016	REG3A	-4.51	3.35e-46	0.003
1	ENSG00000164292	RHOBTB3	-5.19	2.35e-46	0.003
1	ENSG00000167117	LINC00483	-3.58	1.38e-45	0.004
1	ENSG00000087116	ADAMTS2	-3.63	1.91e-45	0.004
1	ENSG00000185885	IFITM1	-3.63	1.45e-45	0.004
1	ENSG00000198133	TMEM229B	-3.87	1.22e-45	0.004
1	ENSG00000261713	SSTR5-AS1	-3.28	1.01e-44	0.006
1	ENSG00000165828	PRAP1	-3.32	1.09e-44	0.006
1	ENSG00000064651	SLC12A2	-3.47	8.35e-45	0.006
1	ENSG00000251095	ENSG00000251095	-3.49	7.14e-45	0.006
1	ENSG00000152932	RAB3C	-3.56	7.57e-45	0.006
1	ENSG00000230316	FEZF1-AS1	-3.86	6.93e-45	0.006
1	ENSG00000206073	SERPINB4	-4.81	1.03e-44	0.006
1	ENSG00000073849	ST6GAL1	-4.84	8.75e-45	0.006
1	ENSG00000167004	PDIA3	-1.91	6.95e-44	0.007
1	ENSG00000128268	MGAT3	-3.17	3.14e-44	0.007
1	ENSG00000159374	M1AP	-3.27	6.56e-44	0.007
1	ENSG00000140105	WARS	-3.41	6.77e-44	0.007
1	ENSG00000106178	CCL24	-3.44	1.01e-43	0.007
1	ENSG00000162989	KCNJ3	-3.49	1.73e-44	0.007
1	ENSG00000101187	SLCO4A1	-3.5	1e-43	0.007
1	ENSG00000196611	MMP1	-3.52	7.34e-44	0.007
1	ENSG00000142789	CELA3A	-3.56	8.13e-44	0.007
1	ENSG00000269189	ENSG00000269189	-3.63	4.11e-44	0.007
1	ENSG00000172238	ATOH1	-3.67	1.09e-43	0.007
1	ENSG00000177707	PVRL3	-3.86	4.7e-44	0.007
1	ENSG00000019505	SYT13	-3.88	7.67e-44	0.007
1	ENSG00000187715	KBTBD12	-3.97	2.21e-44	0.007
1	ENSG00000174992	ZG16	-4.59	6.93e-44	0.007
1	ENSG00000181143	MUC16	-4.68	2.16e-44	0.007
1	ENSG00000196126	HLA-DRB1	-4.87	5.46e-44	0.007
1	ENSG00000167748	KLK1	-5.14	2.52e-44	0.007
1	ENSG00000272620	AFAP1-AS1	-5.16	3.36e-44	0.007

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000215864	NBPF7	-3.48	1.8e-43	0.008
1	ENSG00000271824	ENSG00000271824	-3.52	2.32e-43	0.008
1	ENSG00000178826	TMEM139	-4.43	2.6e-43	0.008
1	ENSG00000154274	C4orf19	-4.49	1.36e-43	0.008
1	ENSG00000174567	GOLT1A	-3.96	2.68e-43	0.009
1	ENSG00000100079	LGALS2	-3.05	5.16e-43	0.01
1	ENSG00000163297	ANTXR2	-4.72	4.4e-43	0.01
1	ENSG00000012223	LTF	-5.58	5.81e-43	0.01
1	ENSG00000163737	PF4	-3.14	1.03e-42	0.012
1	ENSG00000185615	PDIA2	-3.38	7.87e-43	0.012
1	ENSG00000110492	MDK	-3.5	8.03e-43	0.012
1	ENSG00000254290	ENSG00000254290	-3.51	8.79e-43	0.012
1	ENSG00000172159	FRMD3	-3.74	9.01e-43	0.012
1	ENSG00000110324	IL10RA	-3.99	1.12e-42	0.013
1	ENSG00000128422	KRT17	-3.4	1.21e-42	0.014
1	ENSG00000132205	EMILIN2	-3.4	1.23e-42	0.014
1	ENSG00000107249	GLIS3	-4.11	1.18e-42	0.014
1	ENSG00000239893	ZNF736P9Y	-3.06	1.73e-42	0.016
1	ENSG00000020633	RUNX3	-3.66	2.09e-42	0.017
1	ENSG00000228630	HOTAIR	-3.42	2.38e-42	0.018
1	ENSG00000123388	HOXC11	-3.08	3.51e-42	0.019
1	ENSG00000177301	KCNA2	-3.14	3.67e-42	0.019
1	ENSG00000173535	TNFRSF10C	-3.56	5.25e-42	0.021
1	ENSG00000144229	THSD7B	-3.84	7.47e-42	0.021
1	ENSG00000166278	C2	-4.08	4.83e-42	0.021
1	ENSG00000129167	TPH1	-5.76	6.54e-42	0.021
1	ENSG00000116774	OLFML3	-4.11	9.5e-42	0.022
1	ENSG00000130038	CRACR2A	-4.52	1.11e-41	0.022
1	ENSG00000197584	KCNMB2	-3.4	1.56e-41	0.023
1	ENSG00000106003	LFNG	-5.16	1.7e-41	0.024
1	ENSG00000148346	LCN2	-5.23	1.68e-41	0.024
1	ENSG00000170345	FOS	-5.1	2.51e-41	0.026
1	ENSG00000056998	GYG2	-3.12	2.76e-41	0.027
1	ENSG00000186188	FFAR4	-3.38	3.02e-41	0.027
1	ENSG00000269404	SPIB	-3.72	3.94e-41	0.031
1	ENSG00000108176	DNAJC12	-3.21	4.39e-41	0.032
1	ENSG00000170324	FRMPD2	-3.29	4.68e-41	0.033
1	ENSG00000173918	C1QTNF1	-3.71	5.28e-41	0.033
1	ENSG00000179776	CDH5	-4.37	5.95e-41	0.033
1	ENSG00000125999	BPIFB1	-3.72	1.05e-40	0.041
1	ENSG00000196169	KIF19	-3.84	1.29e-40	0.041
1	ENSG00000206159	GYG2P1	-3	1.5e-40	0.042

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000167912	ENSG00000167912	-3	1.46e-40	0.042
1	ENSG00000114115	RBP1	-4.79	1.81e-40	0.046
1	ENSG00000204287	HLA-DRA	-4.89	1.8e-40	0.046
1	ENSG00000157064	NMNAT2	-3.42	2.11e-40	0.048
1	ENSG00000248429	ENSG00000248429	-3.14	2.17e-40	0.049
2	ENSG00000169245	CXCL10	8.96	1.87e-247	0.000
2	ENSG00000138755	CXCL9	8.8	1.57e-320	0.000
2	ENSG00000131203	IDO1	8.24	1.77e-270	0.000
2	ENSG00000173432	SAA1	7.87	6.4e-205	0.000
2	ENSG00000204252	HLA-DOA	7.72	2.06e-180	0.000
2	ENSG00000196735	HLA-DQA1	6.64	1.61e-79	0.000
2	ENSG00000169248	CXCL11	6.54	8.31e-140	0.000
2	ENSG00000154451	GBP5	6.44	4.38e-77	0.000
2	ENSG00000137959	IFI44L	6.32	8.28e-64	0.000
2	ENSG00000143341	HMCN1	6.3	1.7e-60	0.000
2	ENSG00000225630	MTND2P28	5.95	0.000	0.000
2	ENSG00000223865	HLA-DPB1	5.86	6.54e-61	0.000
2	ENSG00000206073	SERPINB4	5.85	1.3e-61	0.000
2	ENSG00000162692	VCAM1	5.66	2.15e-63	0.000
2	ENSG00000205076	LGALS7	5.56	0.000	0.000
2	ENSG00000232629	HLA-DQB2	5.33	1.63e-137	0.000
2	ENSG00000213886	UBD	5.23	1.07e-123	0.000
2	ENSG00000122679	RAMP3	5.06	4.44e-61	0.000
2	ENSG00000229391	HLA-DRB6	5.03	7.55e-116	0.000
2	ENSG00000226025	ENSG00000226025	4.99	3.44e-69	0.000
2	ENSG00000134339	SAA2	4.68	6.77e-85	0.000
2	ENSG00000128422	KRT17	4.42	3.86e-69	0.000
2	ENSG00000204525	HLA-C	2.73	4.34e-86	0.000
2	ENSG00000234745	HLA-B	2.24	7.72e-75	0.000
2	ENSG00000125730	C3	6.41	2.49e-58	0.001
2	ENSG00000198502	HLA-DRB5	6.07	3.31e-57	0.001
2	ENSG00000114115	RBP1	5.9	6.4e-58	0.001
2	ENSG00000108244	KRT23	5.38	6.35e-57	0.002
2	ENSG00000140105	WARS	4.24	1.2e-55	0.003
2	ENSG00000206341	HLA-H	2.62	1.44e-53	0.006
2	ENSG00000231389	HLA-DPA1	6.38	1.72e-51	0.02
2	ENSG00000149131	SERPING1	5.84	6.31e-51	0.028
2	ENSG00000204632	HLA-G	2.66	1.53e-50	0.033
2	ENSG00000213553	RPLP0P6	-1.61	2.14e-59	0.000
2	ENSG00000234797	RPS3AP6	-1.8	5.44e-67	0.000
2	ENSG00000212802	RPL15P3	-1.81	1.24e-61	0.000
2	ENSG00000137970	RPL7P9	-1.92	1.62e-67	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
2	ENSG00000244313	ENSG00000244313	-1.95	1.63e-65	0.000
2	ENSG00000136810	TXN	-1.98	1e-72	0.000
2	ENSG00000213442	RPL18AP3	-1.99	4.9e-113	0.000
2	ENSG00000230202	ENSG00000230202	-2.17	2.45e-59	0.000
2	ENSG00000233762	ENSG00000233762	-2.2	2.52e-74	0.000
2	ENSG00000242299	ENSG00000242299	-2.33	1.24e-112	0.000
2	ENSG00000224631	ENSG00000224631	-2.4	1.52e-65	0.000
2	ENSG00000227081	ENSG00000227081	-2.51	4.05e-93	0.000
2	ENSG00000146677	ENSG00000146677	-2.52	9.91e-65	0.000
2	ENSG00000232573	RPL3P4	-2.52	5.53e-96	0.000
2	ENSG00000214485	RPL7P1	-2.66	8.26e-90	0.000
2	ENSG00000244398	ENSG00000244398	-2.67	1.46e-71	0.000
2	ENSG00000139239	RPL14P1	-2.75	3.94e-87	0.000
2	ENSG00000231767	ENSG00000231767	-2.94	6.65e-60	0.000
2	ENSG00000185641	ENSG00000185641	-2.99	9.24e-96	0.000
2	ENSG00000227615	ENSG00000227615	-3.11	1.08e-68	0.000
2	ENSG00000218426	ENSG00000218426	-3.13	4.42e-147	0.000
2	ENSG00000230629	RPS23P8	-3.19	3.02e-80	0.000
2	ENSG00000213598	ENSG00000213598	-3.28	8.37e-103	0.000
2	ENSG00000226221	RPL26P19	-3.66	1.23e-82	0.000
2	ENSG00000218175	ENSG00000218175	-3.8	5.22e-106	0.000
2	ENSG00000232346	ENSG00000232346	-3.92	2.62e-59	0.000
2	ENSG00000178934	LGALS7B	-4.53	1.01e-81	0.000
2	ENSG00000235552	RPL6P27	-1.78	5.99e-59	0.001
2	ENSG00000187653	TMSB4XP8	-4.05	3.07e-57	0.001
2	ENSG00000236552	RPL13AP5	-1.34	4.14e-56	0.003
2	ENSG00000228929	RPS13P2	-2.83	1.63e-53	0.006
2	ENSG00000229807	XIST	-8.25	9.22e-53	0.008
2	ENSG00000232472	EEF1B2P3	-2.38	1.28e-51	0.02
2	ENSG00000226360	RPL10AP6	-2.06	5.49e-51	0.027
2	ENSG00000226608	FTLP3	-3.85	1.29e-50	0.03
3	ENSG00000215182	MUC5AC	12.1	0.000	0.000
3	ENSG00000184956	MUC6	12	0.000	0.000
3	ENSG00000134193	REG4	11.8	0.000	0.000
3	ENSG00000160182	TFF1	11.3	0.000	0.000
3	ENSG00000102837	OLFM4	11.3	0.000	0.000
3	ENSG00000090382	LYZ	10.9	0.000	0.000
3	ENSG00000197249	SERPINA1	10.9	0.000	0.000
3	ENSG00000160180	TFF3	10.5	0.000	0.000
3	ENSG00000171747	LGALS4	10.1	0.000	0.000
3	ENSG00000162896	PIGR	10.1	1.26e-276	0.000
3	ENSG00000198788	MUC2	10.1	1.44e-146	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000164266	SPINK1	9.93	0.000	0.000
3	ENSG00000187908	DMBT1	9.92	0.000	0.000
3	ENSG00000117983	MUC5B	9.83	0.000	0.000
3	ENSG00000163993	S100P	9.82	0.000	0.000
3	ENSG00000122711	SPINK4	9.72	8.17e-269	0.000
3	ENSG00000196188	CTSE	9.71	0.000	0.000
3	ENSG00000106541	AGR2	9.66	1.68e-151	0.000
3	ENSG00000105388	CEACAM5	9.54	3.18e-156	0.000
3	ENSG00000116299	KIAA1324	9.5	0.000	0.000
3	ENSG00000173467	AGR3	9.13	0.000	0.000
3	ENSG00000109511	ANXA10	9.08	7.74e-269	0.000
3	ENSG00000134398	ERN2	8.92	0.000	0.000
3	ENSG00000157613	CREB3L1	8.92	0.000	0.000
3	ENSG00000198203	SULT1C2	8.86	0.000	0.000
3	ENSG00000173702	MUC13	8.85	0.000	0.000
3	ENSG00000066405	CLDN18	8.77	0.000	0.000
3	ENSG00000115648	MLPH	8.73	0.000	0.000
3	ENSG00000127324	TSPAN8	8.65	4.46e-157	0.000
3	ENSG00000103534	TMC5	8.58	0.000	0.000
3	ENSG00000169876	MUC17	8.51	2.26e-171	0.000
3	ENSG00000064787	BCAS1	8.45	0.000	0.000
3	ENSG00000204385	SLC44A4	8.41	9.35e-257	0.000
3	ENSG00000007062	PROM1	8.37	0.000	0.000
3	ENSG00000179914	ITLN1	8.34	1.94e-109	0.000
3	ENSG00000259974	LINC00261	8.32	7.1e-112	0.000
3	ENSG00000079112	CDH17	8.26	0.000	0.000
3	ENSG00000243709	LEFTY1	8.23	1.57e-145	0.000
3	ENSG00000188175	HEPACAM2	8.23	1.61e-137	0.000
3	ENSG00000165215	CLDN3	8.2	5.48e-310	0.000
3	ENSG00000130513	GDF15	8.16	4.37e-182	0.000
3	ENSG00000129538	RNASE1	8.09	3.2e-211	0.000
3	ENSG00000169894	MUC3A	8.09	8.53e-253	0.000
3	ENSG00000002726	AOC1	8.03	6.81e-288	0.000
3	ENSG00000160181	TFF2	7.97	5.28e-168	0.000
3	ENSG00000127831	VIL1	7.97	3.61e-264	0.000
3	ENSG00000006611	USH1C	7.95	0.000	0.000
3	ENSG00000185499	MUC1	7.87	7.75e-120	0.000
3	ENSG00000140297	GCNT3	7.83	2.51e-138	0.000
3	ENSG00000164816	DEFA5	7.81	4.26e-110	0.000
3	ENSG00000095932	SMIM24	7.79	0.000	0.000
3	ENSG00000214814	FER1L6	7.77	3.33e-216	0.000
3	ENSG00000166959	MS4A8	7.76	1.41e-137	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000137860	SLC28A2	7.73	1.35e-152	0.000
3	ENSG00000124664	SPDEF	7.73	8.62e-193	0.000
3	ENSG00000115386	REG1A	7.72	2.51e-135	0.000
3	ENSG00000233041	PHGR1	7.66	0.000	0.000
3	ENSG00000164822	DEFA6	7.65	6.95e-103	0.000
3	ENSG00000171431	KRT20	7.64	4.79e-117	0.000
3	ENSG00000165140	FBP1	7.61	0.000	0.000
3	ENSG00000010438	PRSS3	7.6	3.58e-164	0.000
3	ENSG00000198758	EPS8L3	7.59	2.52e-214	0.000
3	ENSG00000101443	WFDC2	7.57	8.27e-95	0.000
3	ENSG00000102854	MSLN	7.55	1.16e-224	0.000
3	ENSG00000117525	F3	7.55	3.37e-139	0.000
3	ENSG00000013588	GPRC5A	7.53	4.09e-86	0.000
3	ENSG00000122859	NEUROG3	7.51	2.28e-111	0.000
3	ENSG00000203697	CAPN8	7.51	4.41e-127	0.000
3	ENSG00000143167	GPA33	7.48	2.17e-178	0.000
3	ENSG00000122176	FMOD	7.47	2.73e-159	0.000
3	ENSG00000074370	ATP2A3	7.45	2.61e-155	0.000
3	ENSG00000137501	SYTL2	7.44	2.52e-122	0.000
3	ENSG00000101076	HNF4A	7.41	0.000	0.000
3	ENSG00000163734	CXCL3	7.39	1.17e-142	0.000
3	ENSG00000119547	ONECUT2	7.35	1.21e-171	0.000
3	ENSG00000162545	CAMK2N1	7.33	7.91e-279	0.000
3	ENSG00000267795	SMIM22	7.31	5.70e-318	0.000
3	ENSG00000089472	HEPH	7.3	6.47e-191	0.000
3	ENSG00000176907	C8orf4	7.29	1.94e-211	0.000
3	ENSG00000157005	SST	7.29	5.65e-96	0.000
3	ENSG00000198643	FAM3D	7.29	4.24e-94	0.000
3	ENSG00000169507	SLC38A11	7.27	1.09e-152	0.000
3	ENSG00000103460	TOX3	7.27	1.65e-263	0.000
3	ENSG00000106565	TMEM176B	7.26	2.22e-191	0.000
3	ENSG00000236699	ARHGEF38	7.26	1.19e-290	0.000
3	ENSG00000101333	PLCB4	7.24	1.12e-210	0.000
3	ENSG00000099812	MISP	7.24	2.37e-137	0.000
3	ENSG00000131771	PPP1R1B	7.22	4.33e-89	0.000
3	ENSG00000160862	AZGP1	7.21	1.06e-159	0.000
3	ENSG00000173597	SULT1B1	7.21	1.38e-158	0.000
3	ENSG00000162878	PKDCC	7.2	4.73e-186	0.000
3	ENSG00000169994	MYO7B	7.17	2.05e-170	0.000
3	ENSG00000169213	RAB3B	7.16	1.56e-129	0.000
3	ENSG00000184709	LRRC26	7.14	4.92e-134	0.000
3	ENSG00000088992	TESC	7.12	1.43e-168	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000166922	SCG5	7.12	1.16e-113	0.000
3	ENSG00000158764	ITLN2	7.1	1.03e-78	0.000
3	ENSG00000096088	PGC	7.07	2.78e-155	0.000
3	ENSG00000148735	PLEKHS1	7.07	8.17e-169	0.000
3	ENSG00000101194	SLC17A9	7.07	2.09e-148	0.000
3	ENSG00000106331	PAX4	7.07	4.04e-93	0.000
3	ENSG00000112379	ARFGEF3	7.05	3.75e-97	0.000
3	ENSG00000163898	LIPH	7.04	1.5e-181	0.000
3	ENSG00000120875	DUSP4	7.04	1.69e-211	0.000
3	ENSG00000180745	CLRN3	7.03	6.14e-238	0.000
3	ENSG00000019102	VSIG2	7.01	9.34e-102	0.000
3	ENSG00000141448	GATA6	6.96	0.000	0.000
3	ENSG00000101842	VSIG1	6.94	8.68e-95	0.000
3	ENSG00000168631	DPCR1	6.9	2.86e-100	0.000
3	ENSG00000156966	B3GNT7	6.89	5.35e-125	0.000
3	ENSG00000148584	A1CF	6.88	3.03e-182	0.000
3	ENSG00000178828	RNF186	6.87	1.99e-119	0.000
3	ENSG00000113924	HGD	6.87	9.89e-196	0.000
3	ENSG00000113722	CDX1	6.85	3.16e-159	0.000
3	ENSG00000104783	KCNN4	6.84	7.24e-105	0.000
3	ENSG00000074276	CDHR2	6.82	2.96e-145	0.000
3	ENSG00000123191	ATP7B	6.8	4.92e-135	0.000
3	ENSG00000257743	MGAM2	6.79	5.33e-165	0.000
3	ENSG00000255545	ENSG00000255545	6.76	3.92e-104	0.000
3	ENSG00000070019	GUCY2C	6.74	2.33e-136	0.000
3	ENSG00000169347	GP2	6.72	8.02e-123	0.000
3	ENSG00000096395	MLN	6.72	1.31e-93	0.000
3	ENSG00000198734	F5	6.71	2.14e-157	0.000
3	ENSG00000166825	ANPEP	6.66	2.52e-74	0.000
3	ENSG00000132911	NMUR2	6.61	3.5e-92	0.000
3	ENSG00000183778	B3GALT5	6.6	2.87e-102	0.000
3	ENSG00000198488	B3GNT6	6.59	3.28e-108	0.000
3	ENSG00000132437	DDC	6.57	1.61e-162	0.000
3	ENSG00000250337	LINC01021	6.54	8.72e-135	0.000
3	ENSG00000150594	ADRA2A	6.54	3.2e-161	0.000
3	ENSG00000002933	TMEM176A	6.49	1.32e-111	0.000
3	ENSG00000169562	GJB1	6.48	7.96e-226	0.000
3	ENSG00000142484	TM4SF5	6.47	2.22e-174	0.000
3	ENSG00000184012	TMPRSS2	6.45	2.29e-81	0.000
3	ENSG00000135773	CAPN9	6.42	3.29e-135	0.000
3	ENSG00000165556	CDX2	6.42	4.13e-149	0.000
3	ENSG00000161798	AQP5	6.41	1.68e-139	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000125798	FOXA2	6.41	2.27e-150	0.000
3	ENSG00000163739	CXCL1	6.38	8.33e-124	0.000
3	ENSG00000197635	DPP4	6.36	1.55e-153	0.000
3	ENSG00000174358	SLC6A19	6.36	3.73e-108	0.000
3	ENSG00000170439	METTL7B	6.34	6.3e-113	0.000
3	ENSG00000163586	FABP1	6.34	2.67e-91	0.000
3	ENSG00000053918	KCNQ1	6.33	2.82e-97	0.000
3	ENSG00000196730	DAPK1	6.32	3.74e-130	0.000
3	ENSG00000104537	ANXA13	6.29	8.54e-118	0.000
3	ENSG00000086696	HSD17B2	6.28	4.48e-89	0.000
3	ENSG00000139515	PDX1	6.27	4.46e-135	0.000
3	ENSG00000108511	HOXB6	6.25	3.45e-174	0.000
3	ENSG00000139540	SLC39A5	6.23	5.85e-128	0.000
3	ENSG00000138722	MMRN1	6.2	1.2e-111	0.000
3	ENSG00000148702	HABP2	6.19	2.37e-109	0.000
3	ENSG00000166869	CHP2	6.17	1.01e-88	0.000
3	ENSG00000170608	FOXA3	6.16	5.01e-163	0.000
3	ENSG00000186854	TRABD2A	6.15	2.68e-97	0.000
3	ENSG00000106031	HOXA13	6.13	3.19e-137	0.000
3	ENSG00000204616	TRIM31	6.13	1.57e-100	0.000
3	ENSG00000001626	CFTR	6.09	2.47e-71	0.000
3	ENSG00000129946	SHC2	6.07	2.31e-78	0.000
3	ENSG00000012171	SEMA3B	6.07	6.11e-71	0.000
3	ENSG00000198846	TOX	6.05	9.15e-123	0.000
3	ENSG00000163515	RETNLB	6.05	7.49e-84	0.000
3	ENSG00000166920	C15orf48	6.05	4.61e-89	0.000
3	ENSG00000204128	C2orf72	6.04	4.16e-80	0.000
3	ENSG00000107242	PIP5K1B	6.02	2.27e-145	0.000
3	ENSG00000123358	NR4A1	6.01	7.68e-78	0.000
3	ENSG00000183018	SPNS2	5.99	5.08e-145	0.000
3	ENSG00000021300	PLEKHB1	5.99	4.56e-90	0.000
3	ENSG00000100170	SLC5A1	5.98	1.77e-107	0.000
3	ENSG00000100302	RASD2	5.97	1.35e-88	0.000
3	ENSG00000204876	ENSG00000204876	5.97	1.85e-107	0.000
3	ENSG00000009765	IYD	5.95	3.6e-120	0.000
3	ENSG00000166866	MYO1A	5.95	5.94e-115	0.000
3	ENSG00000088340	FER1L4	5.93	4.97e-73	0.000
3	ENSG00000166510	CCDC68	5.92	1.7e-89	0.000
3	ENSG00000204019	CT83	5.89	1.19e-141	0.000
3	ENSG00000115616	SLC9A2	5.88	7.19e-126	0.000
3	ENSG00000179674	ARL14	5.85	2.46e-107	0.000
3	ENSG00000101577	LPIN2	5.84	3.96e-78	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000163072	NOSTRIN	5.84	8.37e-97	0.000
3	ENSG00000261857	MIA	5.81	4.34e-122	0.000
3	ENSG00000162174	ASRGL1	5.8	2.35e-96	0.000
3	ENSG00000113303	BTNL8	5.8	3.53e-119	0.000
3	ENSG00000243766	HOTTIP	5.78	5.31e-87	0.000
3	ENSG00000131910	NR0B2	5.78	3.24e-91	0.000
3	ENSG00000120075	HOXB5	5.76	6.64e-116	0.000
3	ENSG00000136574	GATA4	5.75	3.61e-101	0.000
3	ENSG00000153551	CMTM7	5.74	2.42e-96	0.000
3	ENSG00000165507	C10orf10	5.72	1.53e-78	0.000
3	ENSG00000178882	FAM101A	5.68	4.42e-105	0.000
3	ENSG00000099834	CDHR5	5.64	6.06e-110	0.000
3	ENSG00000188833	ENTPD8	5.64	3.25e-76	0.000
3	ENSG00000188993	LRRC66	5.63	1.04e-79	0.000
3	ENSG00000123843	C4BPB	5.62	1.7e-106	0.000
3	ENSG00000139292	LGR5	5.61	1.33e-77	0.000
3	ENSG00000118017	A4GNT	5.57	4.09e-78	0.000
3	ENSG00000204610	TRIM15	5.55	6.39e-114	0.000
3	ENSG00000248771	LINC01207	5.54	1.39e-72	0.000
3	ENSG00000106069	CHN2	5.53	1.08e-84	0.000
3	ENSG00000015520	NPC1L1	5.52	9.8e-77	0.000
3	ENSG00000172016	REG3A	5.52	1.1e-73	0.000
3	ENSG00000111110	PPM1H	5.49	2.17e-117	0.000
3	ENSG00000175311	ANKS4B	5.42	1.19e-80	0.000
3	ENSG00000248596	ENSG00000248596	5.39	2.93e-77	0.000
3	ENSG00000140968	IRF8	5.39	2.58e-99	0.000
3	ENSG00000145384	FABP2	5.38	3.37e-91	0.000
3	ENSG00000159184	HOXB13	5.36	5.29e-99	0.000
3	ENSG00000101670	LIPG	5.31	2.69e-73	0.000
3	ENSG00000240990	HOXA11-AS	5.3	1.13e-90	0.000
3	ENSG00000163501	IHH	5.28	1.65e-104	0.000
3	ENSG00000198535	C2CD4A	5.23	2.68e-100	0.000
3	ENSG00000180251	SLC9A4	5.22	2.44e-82	0.000
3	ENSG00000169903	TM4SF4	5.16	4.41e-72	0.000
3	ENSG00000130035	GALNT8	5.13	4.65e-74	0.000
3	ENSG00000125384	PTGER2	5.13	2.23e-87	0.000
3	ENSG00000016082	ISL1	5.03	2.43e-75	0.000
3	ENSG00000266010	GATA6-AS1	5.02	3.77e-83	0.000
3	ENSG00000146039	SLC17A4	4.96	2.11e-71	0.000
3	ENSG00000168903	BTNL3	4.96	2.5e-80	0.000
3	ENSG00000165626	BEND7	4.96	8.37e-77	0.000
3	ENSG00000114248	LRRC31	4.93	5.99e-100	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000224081	LINC01057	4.93	3.6e-84	0.000
3	ENSG00000090402	SI	4.9	1.43e-84	0.000
3	ENSG00000157399	ARSE	4.86	1.78e-88	0.000
3	ENSG00000166143	PPP1R14D	4.85	5.02e-77	0.000
3	ENSG00000099617	EFNA2	4.81	4e-97	0.000
3	ENSG00000005073	HOXA11	4.8	1.08e-84	0.000
3	ENSG00000205922	ONECUT3	4.8	2.46e-90	0.000
3	ENSG00000137251	TINAG	4.74	1.29e-75	0.000
3	ENSG00000105251	SHD	4.7	3.51e-79	0.000
3	ENSG00000138100	TRIM54	4.7	8.21e-89	0.000
3	ENSG00000182333	LIPF	4.59	4.04e-73	0.000
3	ENSG00000146038	DCDC2	4.59	1.58e-79	0.000
3	ENSG00000111701	APOBEC1	4.51	1.5e-77	0.000
3	ENSG00000162779	AXDND1	4.47	9.41e-81	0.000
3	ENSG00000272763	ENSG00000272763	4.41	1.1e-78	0.000
3	ENSG00000267541	MTCO2P2	3.98	5.68e-99	0.000
3	ENSG00000270225	ENSG00000270225	3.83	1.68e-73	0.000
3	ENSG00000229031	ENSG00000229031	3.83	4.94e-81	0.000
3	ENSG00000197956	S100A6	1.89	1.55e-74	0.000
3	ENSG00000135480	KRT7	6.73	2.13e-67	0.001
3	ENSG00000135052	GOLM1	6.67	1.14e-67	0.001
3	ENSG00000081041	CXCL2	6.65	9.1e-68	0.001
3	ENSG00000108551	RASD1	6.24	1.11e-67	0.001
3	ENSG00000272620	AFAP1-AS1	6.2	1.17e-64	0.001
3	ENSG00000167748	KLK1	6.19	1.08e-64	0.001
3	ENSG00000172482	AGXT	6.19	2.64e-67	0.001
3	ENSG00000117707	PROX1	6.15	6.05e-64	0.001
3	ENSG00000100604	CHGA	5.98	6.02e-65	0.001
3	ENSG00000160183	TMPRSS3	5.96	8.67e-66	0.001
3	ENSG00000112699	GMDS	5.95	1.42e-69	0.001
3	ENSG00000162078	ZG16B	5.86	8.86e-68	0.001
3	ENSG00000163297	ANTXR2	5.8	8.45e-71	0.001
3	ENSG00000142185	TRPM2	5.78	3.57e-64	0.001
3	ENSG00000174992	ZG16	5.6	1.29e-63	0.001
3	ENSG00000196436	NPIPB15	5.54	6.47e-66	0.001
3	ENSG00000154274	C4orf19	5.52	6.98e-68	0.001
3	ENSG00000105707	HPN	5.49	8.94e-65	0.001
3	ENSG00000168955	TM4SF20	5.49	4.14e-63	0.001
3	ENSG00000106772	PRUNE2	5.47	2.16e-64	0.001
3	ENSG00000163817	SLC6A20	5.44	5.64e-63	0.001
3	ENSG00000160188	RSPH1	5.38	1.87e-64	0.001
3	ENSG00000128917	DLL4	5.34	1.36e-69	0.001

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000183960	KCNH8	5.27	5.48e-68	0.001
3	ENSG00000112818	MEP1A	5.19	2.39e-70	0.001
3	ENSG00000107249	GLIS3	5.14	2.93e-69	0.001
3	ENSG00000187715	KBTBD12	4.97	2.83e-69	0.001
3	ENSG00000174567	GOLT1A	4.96	3.19e-68	0.001
3	ENSG00000173237	C11orf86	4.92	1.44e-70	0.001
3	ENSG00000019505	SYT13	4.86	1.02e-65	0.001
3	ENSG00000230316	FEZF1-AS1	4.85	1.27e-68	0.001
3	ENSG00000114455	HHLA2	4.76	1.12e-69	0.001
3	ENSG00000150750	C11orf53	4.73	1.31e-70	0.001
3	ENSG00000138207	RBP4	4.67	8.65e-67	0.001
3	ENSG00000130957	FBP2	4.63	9.14e-67	0.001
3	ENSG00000213996	TM6SF2	4.6	5.39e-66	0.001
3	ENSG00000167117	LINC00483	4.56	5.36e-70	0.001
3	ENSG00000251350	ENSG00000251350	4.54	7.21e-69	0.001
3	ENSG00000271824	ENSG00000271824	4.5	2.57e-63	0.001
3	ENSG00000162989	KCNJ3	4.47	9.89e-66	0.001
3	ENSG00000251095	ENSG00000251095	4.46	3.26e-63	0.001
3	ENSG00000165828	PRAP1	4.29	1.61e-67	0.001
3	ENSG00000261713	SSTR5-AS1	4.25	1.31e-64	0.001
3	ENSG00000205542	TMSB4X	2	4.44e-70	0.001
3	ENSG00000179913	B3GNT3	6.22	1.75e-61	0.002
3	ENSG00000106003	LFNG	6.21	3.42e-61	0.002
3	ENSG00000177685	CRACR2B	6.07	4.55e-60	0.002
3	ENSG00000165376	CLDN2	5.92	1.61e-61	0.002
3	ENSG00000250069	ENSG00000250069	5.88	2.27e-61	0.002
3	ENSG00000183508	FAM46C	5.84	1.22e-58	0.002
3	ENSG00000131941	RHPN2	5.53	1.55e-59	0.002
3	ENSG00000179776	CDH5	5.4	1.98e-58	0.002
3	ENSG00000166391	MOGAT2	5.34	4.66e-60	0.002
3	ENSG00000136052	SLC41A2	5.25	1.27e-59	0.002
3	ENSG00000095539	SEMA4G	5.18	1.95e-58	0.002
3	ENSG00000198286	CARD11	4.77	2.27e-59	0.002
3	ENSG00000172238	ATOH1	4.65	4.05e-62	0.002
3	ENSG00000269189	ENSG00000269189	4.62	9.61e-61	0.002
3	ENSG00000142789	CELA3A	4.51	1.83e-60	0.002
3	ENSG00000152932	RAB3C	4.48	1.4e-58	0.002
3	ENSG00000254290	ENSG00000254290	4.46	4.82e-59	0.002
3	ENSG00000215864	NBPF7	4.46	3.66e-60	0.002
3	ENSG00000101187	SLCO4A1	4.45	1.94e-60	0.002
3	ENSG00000106178	CCL24	4.43	1.86e-61	0.002
3	ENSG00000228630	HOTAIR	4.39	1e-60	0.002

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000259120	SMIM6	4.39	7.12e-59	0.002
3	ENSG00000170703	TTLL6	4.32	4.49e-60	0.002
3	ENSG00000163737	PF4	4.09	1.96e-61	0.002
3	ENSG00000128268	MGAT3	4.07	2.67e-61	0.002
3	ENSG00000100079	LGALS2	4.02	1.65e-62	0.002
3	ENSG00000111057	KRT18	3.31	3.04e-59	0.002
3	ENSG00000140264	SERF2	1.23	2.13e-58	0.002
3	ENSG00000129167	TPH1	6.8	2.44e-57	0.003
3	ENSG00000145113	MUC4	6.39	1.55e-57	0.003
3	ENSG00000176485	PLA2G16	6.04	4.37e-58	0.003
3	ENSG00000009950	MLXIPL	5.97	3.88e-57	0.003
3	ENSG00000205502	C2CD4B	5.29	4.44e-58	0.003
3	ENSG00000126218	F10	4.99	2.23e-57	0.003
3	ENSG00000004848	ARX	4.8	2.89e-57	0.003
3	ENSG00000269404	SPIB	4.67	2.23e-57	0.003
3	ENSG00000073464	CLCN4	4.53	1.34e-57	0.003
3	ENSG00000164690	SHH	4.48	2.04e-56	0.003
3	ENSG00000197584	KCNMB2	4.36	5.09e-57	0.003
3	ENSG00000157064	NMNAT2	4.36	1.49e-56	0.003
3	ENSG00000186188	FFAR4	4.33	1.21e-57	0.003
3	ENSG00000185615	PDIA2	4.33	3.72e-57	0.003
3	ENSG00000170324	FRMPD2	4.24	6.82e-57	0.003
3	ENSG00000167912	ENSG00000167912	3.96	5.28e-58	0.003
3	ENSG00000241224	ENSG00000241224	3.94	6.33e-58	0.003
3	ENSG00000267015	ENSG00000267015	3.72	2.15e-57	0.003
3	ENSG00000116774	OLFML3	5.08	3.79e-56	0.004
3	ENSG00000196611	MMP1	4.42	6.79e-56	0.004
3	ENSG00000123388	HOXC11	3.97	2.51e-55	0.006
3	ENSG00000153071	DAB2	3.96	4.37e-55	0.007
3	ENSG00000167861	HID1	6.09	2.16e-54	0.009
3	ENSG00000099998	GGT5	5.48	4.44e-54	0.009
3	ENSG00000178401	DNAJC22	5.22	7.52e-55	0.009
3	ENSG00000196169	KIF19	4.8	9.68e-55	0.009
3	ENSG00000007171	NOS2	4.71	1.27e-54	0.009
3	ENSG00000173557	C2orf70	4.37	9e-55	0.009
3	ENSG00000064651	SLC12A2	4.22	9.25e-55	0.009
3	ENSG00000241388	HNF1A-AS1	3.85	2.61e-54	0.009
3	ENSG00000135100	HNF1A	4.86	1.05e-53	0.01
3	ENSG00000227053	ENSG00000227053	3.72	6.77e-54	0.01
3	ENSG00000205517	RGL3	4.45	2.9e-53	0.012
3	ENSG00000214039	ENSG00000214039	4.23	3.28e-53	0.012
3	ENSG00000240024	LINC00888	4.17	3.34e-53	0.012

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000163618	CADPS	7.15	4.77e-53	0.013
3	ENSG00000080493	SLC4A4	5.3	1.28e-52	0.015
3	ENSG00000118271	TTR	4.31	1.23e-52	0.015
3	ENSG00000115423	DNAH6	4.18	9.84e-53	0.015
3	ENSG00000107159	CA9	5.96	2.06e-52	0.016
3	ENSG00000144852	NR1I2	4.45	1.79e-52	0.016
3	ENSG00000010810	FYN	4.01	1.67e-52	0.016
3	ENSG00000125999	BPIFB1	4.63	6.87e-52	0.026
3	ENSG00000182557	SPNS3	4.15	7.88e-52	0.03
3	ENSG00000196684	HSH2D	5.14	9.7e-52	0.031
3	ENSG00000198842	DUSP27	4.28	1.34e-51	0.034
3	ENSG00000173918	C1QTNF1	4.58	2.25e-51	0.037
3	ENSG00000178538	CA8	5.24	2.33e-51	0.038
3	ENSG00000119888	EPCAM	4.66	2.63e-51	0.038
3	ENSG00000169218	RSPO1	3.34	2.88e-51	0.041
3	ENSG00000187889	C1orf168	5.27	3.47e-51	0.042
3	ENSG00000238133	MLK7-AS1	4.78	3.6e-51	0.042
3	ENSG00000231826	ENSG00000231826	4.08	3.89e-51	0.042
3	ENSG00000138771	SHROOM3	5.61	5.44e-51	0.044
3	ENSG00000197273	GUCA2A	4.47	6.1e-51	0.045
3	ENSG00000143768	LEFTY2	3.51	5.83e-51	0.045
3	ENSG00000164687	FABP5	-2.87	1.99e-91	0.000
3	ENSG00000234964	FABP5P7	-3.28	4.34e-89	0.000
3	ENSG00000112378	PERP	-3.42	9.85e-130	0.000
3	ENSG00000175793	SFN	-3.63	2.03e-83	0.000
3	ENSG00000135046	ANXA1	-4.45	1.13e-86	0.000
3	ENSG00000106211	HSPB1	-4.47	1.06e-213	0.000
3	ENSG00000102007	PLP2	-4.59	2.13e-99	0.000
3	ENSG00000165272	AQP3	-5.08	2.95e-78	0.000
3	ENSG00000171346	KRT15	-5.43	4.25e-86	0.000
3	ENSG00000134757	DSG3	-6.41	1.69e-128	0.000
3	ENSG00000137699	TRIM29	-7.14	1.35e-136	0.000
3	ENSG00000196754	S100A2	-7.59	0.000	0.000
3	ENSG00000178934	LGALS7B	-7.76	1.24e-146	0.000
3	ENSG00000121552	CSTA	-7.95	3.18e-150	0.000
3	ENSG00000143320	CRABP2	-7.99	1.76e-151	0.000
3	ENSG00000105894	PTN	-8.34	3.16e-85	0.000
3	ENSG00000163331	DAPL1	-8.34	3.98e-133	0.000
3	ENSG00000137975	CLCA2	-8.39	3.31e-118	0.000
3	ENSG00000081277	PKP1	-8.83	1.87e-97	0.000
3	ENSG00000134762	DSC3	-8.92	1.46e-167	0.000
3	ENSG00000230937	MIR205HG	-9.31	1.75e-184	0.000

Table A.9 Esophageal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000178363	CALML3	-9.32	1.27e-167	0.000
3	ENSG00000124466	LYPD3	-9.51	4.09e-111	0.000
3	ENSG00000183347	GBP6	-9.86	2.61e-120	0.000
3	ENSG00000163220	S100A9	-9.96	1.46e-156	0.000
3	ENSG00000196344	ADH7	-10	4.59e-113	0.000
3	ENSG00000143546	S100A8	-10.1	5.32e-103	0.000
3	ENSG00000167656	LY6D	-10.1	0.000	0.000
3	ENSG00000140519	RHCG	-10.2	1e-101	0.000
3	ENSG00000205420	KRT6A	-10.3	1.13e-111	0.000
3	ENSG00000171401	KRT13	-11.1	0.000	0.000
3	ENSG00000186081	KRT5	-13.1	0.000	0.000
3	ENSG00000187193	MT1X	-2.56	1.41e-68	0.001
3	ENSG00000073282	TP63	-8.39	8.45e-71	0.001
3	ENSG00000175183	CSRP2	-6.04	3.5e-61	0.002
3	ENSG00000256812	CAPNS2	-8.68	8.49e-60	0.002
3	ENSG00000010278	CD9	-2.2	8.24e-51	0.047

Table A.9: Esophageal differentially expressed genes. The table displays the genes identified as significant (FDR <0.05) in differential expression analysis of the gastric single cell clusters. Genes are ordered by cluster and significance and the log2 fold-change in the cluster compared to the rest of single cell libraries is listed for each gene. The average expression across all considered single cell libraries in log2 counts per million is also listed.

A.2.4 Differentially expressed genes across the gastric clusters

Table A.10 Gastric differentially expressed genes.

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000012223	LTF	7.96	7.39e-17	0.000
1	ENSG00000105641	SLC5A5	7.47	2.03e-18	0.000
1	ENSG00000145113	MUC4	6.62	7.08e-28	0.000
1	ENSG00000231389	HLA-DPA1	5.92	4.09e-45	0.000
1	ENSG00000148346	LCN2	5.37	3.88e-40	0.000
1	ENSG00000229391	HLA-DRB6	5.2	8.58e-79	0.000
1	ENSG00000223865	HLA-DPB1	4.43	3.95e-21	0.000
1	ENSG00000179344	HLA-DQB1	4.21	5.34e-21	0.000
1	ENSG00000188818	ZDHHC11	4.16	1.64e-13	0.000
1	ENSG00000204287	HLA-DRA	3.5	1.23e-19	0.000
1	ENSG00000196126	HLA-DRB1	3.5	8.73e-24	0.000
1	ENSG00000198502	HLA-DRB5	3.46	1.13e-22	0.000
1	ENSG00000198643	FAM3D	3.32	5.29e-19	0.000
1	ENSG00000174501	ANKRD36C	3.18	1.96e-19	0.000
1	ENSG00000176153	GPX2	3.07	1.77e-13	0.000
1	ENSG00000162896	PIGR	3.07	3.95e-37	0.000
1	ENSG00000019582	CD74	2.95	1.9e-28	0.000
1	ENSG00000165949	IFI27	2.41	9.43e-15	0.000
1	ENSG00000090382	LYZ	2.09	3.48e-16	0.000
1	ENSG00000234745	HLA-B	1.56	4.65e-14	0.000
1	ENSG00000143248	RGS5	5.57	9.84e-13	0.003
1	ENSG00000100911	PSME2	2.01	2.34e-12	0.007
1	ENSG00000023445	BIRC3	4.11	5.12e-12	0.008
1	ENSG00000204642	HLA-F	2.8	4.95e-12	0.008
1	ENSG00000163131	CTSS	2.28	3.38e-12	0.008
1	ENSG00000110446	SLC15A3	4.05	1.97e-11	0.037
1	ENSG00000170345	FOS	3.27	2.8e-11	0.049
1	ENSG00000225630	MTND2P28	-4.35	1.04e-39	0.000
1	ENSG00000196436	NPIPB15	-4.59	1.85e-14	0.000
1	ENSG00000131002	TXLNGY	-4.99	2.28e-17	0.000
1	ENSG00000133958	UNC79	-5.13	2.94e-13	0.000
1	ENSG00000104112	SCG3	-5.15	8.56e-15	0.000
1	ENSG00000132821	VSTM2L	-5.24	1.43e-13	0.000
1	ENSG00000167964	RAB26	-5.46	2.02e-15	0.000
1	ENSG00000163631	ALB	-5.48	3.15e-20	0.000
1	ENSG00000102466	FGF14	-5.49	4.87e-16	0.000
1	ENSG00000117707	PROX1	-5.49	3.85e-15	0.000
1	ENSG00000176956	LY6H	-5.68	1.42e-14	0.000

Table A.10 Gastric differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000175426	PCSK1	-5.7	2.89e-15	0.000
1	ENSG00000177669	MBOAT4	-5.71	1.7e-13	0.000
1	ENSG00000165507	C10orf10	-5.75	5.82e-14	0.000
1	ENSG00000185002	RFX6	-5.83	1.76e-15	0.000
1	ENSG00000060709	RIMBP2	-5.9	7.37e-18	0.000
1	ENSG00000133169	BEX1	-5.91	1.19e-16	0.000
1	ENSG00000137860	SLC28A2	-6.12	1.19e-21	0.000
1	ENSG00000136698	CFC1	-6.15	1.92e-14	0.000
1	ENSG00000089199	CHGB	-6.18	1.07e-14	0.000
1	ENSG00000132639	SNAP25	-6.23	4e-16	0.000
1	ENSG00000145198	VWA5B2	-6.34	7.5e-16	0.000
1	ENSG00000069535	MAOB	-6.39	9.22e-14	0.000
1	ENSG00000188175	HEPACAM2	-6.42	1.21e-15	0.000
1	ENSG00000162992	NEUROD1	-6.48	6.29e-17	0.000
1	ENSG00000109072	VTN	-6.66	1.32e-14	0.000
1	ENSG00000129824	RPS4Y1	-6.73	4.35e-30	0.000
1	ENSG00000118271	TTR	-6.81	5.01e-14	0.000
1	ENSG00000109472	CPE	-6.9	1.11e-19	0.000
1	ENSG00000102109	PCSK1N	-6.94	1.91e-19	0.000
1	ENSG00000079689	SCGN	-6.97	1.68e-18	0.000
1	ENSG00000157017	GHRL	-7.04	6.19e-30	0.000
1	ENSG00000140287	HDC	-8.2	4.41e-14	0.000
1	ENSG00000100604	CHGA	-8.77	1.69e-26	0.000
1	ENSG00000233864	TTTY15	-3.8	7.11e-13	0.001
1	ENSG00000136842	TMOD1	-4.79	7.02e-13	0.001
1	ENSG00000004848	ARX	-4.8	5.8e-13	0.001
1	ENSG00000012817	KDM5D	-4.85	5.99e-13	0.001
1	ENSG00000157005	SST	-7.23	8.65e-13	0.002
1	ENSG00000184702	SEPT5	-4.69	1.63e-12	0.004
1	ENSG00000006071	ABCC8	-4.97	1.39e-12	0.004
1	ENSG00000174080	CTSF	-4.5	2.16e-12	0.007
1	ENSG00000138722	MMRN1	-5.27	1.8e-12	0.007
1	ENSG00000170091	ENSG00000170091	-4.7	3.44e-12	0.008
1	ENSG00000169213	RAB3B	-4.73	6.95e-12	0.013
1	ENSG00000148798	INA	-4.97	1.1e-11	0.019
1	ENSG00000147488	ST18	-4.57	1.27e-11	0.021
1	ENSG00000129167	TPH1	-6.06	1.34e-11	0.022
1	ENSG00000151726	ACSL1	-4.59	2.25e-11	0.041
1	ENSG00000139767	SRRM4	-4.88	2.27e-11	0.041
2	ENSG00000137860	SLC28A2	6.52	7.41e-33	0.000
2	ENSG00000163631	ALB	5.96	3e-32	0.000
2	ENSG00000131002	TXLNGY	4.98	1.2e-18	0.000

Table A.10 Gastric differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
2	ENSG00000225630	MTND2P28	3.54	2.75e-27	0.000
2	ENSG00000152785	BMP3	4.32	1.42e-14	0.019
2	ENSG00000229314	ORM1	3.89	1.59e-13	0.047
2	ENSG00000166710	B2M	-1.64	4.47e-19	0.000
2	ENSG00000234745	HLA-B	-2.03	7.9e-21	0.000
2	ENSG00000206503	HLA-A	-2.16	1.75e-27	0.000
2	ENSG00000162896	PIGR	-2.51	1.23e-17	0.000
2	ENSG00000204622	HLA-J	-3.17	2.73e-20	0.000
2	ENSG00000229391	HLA-DRB6	-4.83	3.08e-40	0.000
2	ENSG00000148346	LCN2	-4.86	1.26e-21	0.000
2	ENSG00000231389	HLA-DPA1	-5.49	1.55e-24	0.000
2	ENSG00000145113	MUC4	-6.52	5.19e-19	0.000
2	ENSG00000019582	CD74	-2.55	3.68e-16	0.001
2	ENSG00000174501	ANKRD36C	-3.13	7.27e-15	0.011
3	ENSG00000140287	HDC	10	1.66e-36	0.000
3	ENSG00000100604	CHGA	9.42	1.62e-72	0.000
3	ENSG00000118271	TTR	9.14	2.01e-76	0.000
3	ENSG00000109072	VTN	8.99	2.22e-94	0.000
3	ENSG00000079689	SCGN	8.96	1.54e-141	0.000
3	ENSG00000102109	PCSK1N	8.88	2.17e-172	0.000
3	ENSG00000109472	CPE	8.81	2.21e-152	0.000
3	ENSG00000188175	HEPACAM2	8.77	3.14e-139	0.000
3	ENSG00000162992	NEUROD1	8.48	4.93e-96	0.000
3	ENSG00000145198	VWA5B2	8.37	2.68e-74	0.000
3	ENSG00000089199	CHGB	8.31	8.36e-63	0.000
3	ENSG00000132639	SNAP25	8.31	6.79e-89	0.000
3	ENSG00000136698	CFC1	8.3	1.09e-78	0.000
3	ENSG00000054356	PTPRN	8.21	9.31e-62	0.000
3	ENSG00000133169	BEX1	7.98	1.06e-112	0.000
3	ENSG00000185002	RFX6	7.96	4.99e-90	0.000
3	ENSG00000176956	LY6H	7.81	2.14e-108	0.000
3	ENSG00000157017	GHRL	7.81	1.64e-61	0.000
3	ENSG00000114115	RBP1	7.79	2.85e-54	0.000
3	ENSG00000177669	MBOAT4	7.76	5.87e-49	0.000
3	ENSG00000117707	PROX1	7.54	6.1e-89	0.000
3	ENSG00000132821	VSTM2L	7.43	2.3e-90	0.000
3	ENSG00000138722	MMRN1	7.33	2.84e-43	0.000
3	ENSG00000133958	UNC79	7.27	7.55e-71	0.000
3	ENSG00000104112	SCG3	7.19	3.82e-80	0.000
3	ENSG00000139767	SRRM4	7.13	1.74e-60	0.000
3	ENSG00000148798	INA	7.06	1.07e-42	0.000
3	ENSG00000151726	ACSL1	6.83	1.86e-34	0.000

Table A.10 Gastric differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000170091	ENSG00000170091	6.81	7.28e-49	0.000
3	ENSG00000169213	RAB3B	6.73	2.05e-44	0.000
3	ENSG00000145920	CPLX2	6.73	3.81e-47	0.000
3	ENSG00000006071	ABCC8	6.72	2.82e-34	0.000
3	ENSG00000004848	ARX	6.66	3.11e-49	0.000
3	ENSG00000178695	KCTD12	6.64	7.07e-40	0.000
3	ENSG00000147488	ST18	6.54	1.47e-41	0.000
3	ENSG00000126218	F10	6.51	3.03e-36	0.000
3	ENSG00000176463	SLCO3A1	6.51	2.6e-34	0.000
3	ENSG00000091428	RAPGEF4	6.46	1.03e-34	0.000
3	ENSG00000132872	SYT4	6.23	1.44e-43	0.000
3	ENSG00000159409	CELF3	6.05	6.29e-50	0.000
3	ENSG00000124140	SLC12A5	5.97	4.83e-36	0.000
3	ENSG00000125820	NKX2-2	5.83	8.61e-40	0.000
3	ENSG00000165507	C10orf10	7.22	2.15e-30	0.001
3	ENSG00000139219	COL2A1	7.03	7.03e-29	0.001
3	ENSG00000175426	PCSK1	6.79	7.7e-28	0.001
3	ENSG00000060709	RIMBP2	6.79	1.03e-33	0.001
3	ENSG00000007516	BAIAP3	6.75	4.66e-27	0.001
3	ENSG00000196834	POTEI	6.68	8.01e-28	0.001
3	ENSG00000162552	WNT4	6.65	1.48e-33	0.001
3	ENSG00000178162	FAR2P2	6.63	2.74e-28	0.001
3	ENSG00000058404	CAMK2B	6.62	1.42e-29	0.001
3	ENSG00000148584	A1CF	6.37	4.41e-27	0.001
3	ENSG00000136842	TMOD1	6.31	5.46e-27	0.001
3	ENSG00000163618	CADPS	6.29	6.25e-28	0.001
3	ENSG00000138356	AOX1	6.15	1.78e-29	0.001
3	ENSG00000173404	INSM1	6.09	1.46e-28	0.001
3	ENSG00000198732	SMOC1	5.98	9.04e-29	0.001
3	ENSG00000183023	SLC8A1	5.88	4.37e-27	0.001
3	ENSG00000156113	KCNMA1	5.78	6.15e-33	0.001
3	ENSG00000131711	MAP1B	5.7	2.15e-29	0.001
3	ENSG00000087258	GNAO1	5.51	1.6e-26	0.001
3	ENSG00000184905	TCEAL2	5.48	1.85e-26	0.001
3	ENSG00000176406	RIMS2	5.41	1.48e-26	0.001
3	ENSG00000168772	CXXC4	5.24	6.41e-28	0.001
3	ENSG00000267594	CYP4F24P	4.7	1.6e-29	0.001
3	ENSG00000011347	SYT7	5.51	4.7e-26	0.002
3	ENSG00000168243	GNG4	6.09	2.44e-25	0.004
3	ENSG00000104435	STMN2	5.85	1.6e-25	0.004
3	ENSG00000107954	NEURL1	5.47	2.97e-25	0.005
3	ENSG00000102466	FGF14	6.42	1.36e-24	0.008

Table A.10 Gastric differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000043591	ADRB1	5.25	5.75e-24	0.026
3	ENSG00000157005	SST	8.43	6.39e-24	0.028
3	ENSG00000116774	OLFML3	7	6.35e-24	0.028
3	ENSG00000152932	RAB3C	5.62	9.3e-24	0.034
3	ENSG00000172478	C2orf54	5.56	9.56e-24	0.034
3	ENSG00000166741	NNMT	6.87	1.8e-23	0.045
3	ENSG00000205542	TMSB4X	-4.07	1.01e-41	0.000
3	ENSG00000196188	CTSE	-8.23	2.08e-39	0.000
3	ENSG00000106541	AGR2	-4.36	5.99e-28	0.001
3	ENSG00000248144	ADH1C	-7.03	1.29e-26	0.001
3	ENSG00000215182	MUC5AC	-7.13	1.69e-26	0.001
3	ENSG00000163993	S100P	-7.72	1.84e-31	0.001
3	ENSG00000189377	CXCL17	-7.79	9.6e-32	0.001
3	ENSG00000107159	CA9	-8.12	1.2e-28	0.001
3	ENSG00000116260	QSOX1	-8.35	3.36e-31	0.001
3	ENSG00000108602	ALDH3A1	-7.74	6.06e-26	0.002
3	ENSG00000103534	TMC5	-7.1	1.11e-25	0.004

Table A.10: Gastric differentially expressed genes. The table displays the genes identified as significant (FDR <0.05) in differential expression analysis of the gastric single cell clusters. Genes are ordered by cluster and significance and the log2 fold-change in the cluster compared to the rest of gastric single cell libraries is listed for each gene. The average expression across across all considered single cell libraries in log2 counts per million is also listed.

A.2.5 Differentially expressed genes across the duodenal clusters

Table A.9 Duodenal differentially expressed genes.

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000132465	JCHAIN	11.3	6.58e-65	0.000
1	ENSG00000170476	MZB1	10.9	1.26e-42	0.000
1	ENSG00000211949	IGHV3-23	10	7.41e-40	0.000
1	ENSG00000211890	IGHA2	9.86	1.72e-31	0.000
1	ENSG00000026751	SLAMF7	9.6	1.27e-64	0.000
1	ENSG00000230006	ANKRD36BP2	9.34	1.62e-51	0.000
1	ENSG00000048462	TNFRSF17	8.9	9.3e-69	0.000
1	ENSG00000211592	IGKC	8.69	8.5e-58	0.000
1	ENSG00000137265	IRF4	8.59	9.31e-28	0.000
1	ENSG00000211938	IGHV3-7	8.55	1.93e-43	0.000
1	ENSG00000116748	AMPD1	8.47	1.03e-38	0.000
1	ENSG00000211950	IGHV1-24	8.45	1.66e-25	0.000
1	ENSG00000211677	IGLC2	8.43	1.47e-46	0.000
1	ENSG00000167077	MEI1	8.32	2.08e-45	0.000
1	ENSG00000122862	SRGN	8.29	5.3e-32	0.000
1	ENSG00000167476	JSRP1	8.09	8.76e-31	0.000
1	ENSG00000139193	CD27	7.97	8.09e-31	0.000
1	ENSG00000185862	EVI2B	7.92	8.04e-42	0.000
1	ENSG00000102265	TIMP1	7.86	5.09e-42	0.000
1	ENSG00000105369	CD79A	7.75	2.65e-33	0.000
1	ENSG00000211679	IGLC3	7.59	9.44e-27	0.000
1	ENSG00000164932	CTHRC1	7.37	2.26e-28	0.000
1	ENSG00000026025	VIM	7.36	6.49e-30	0.000
1	ENSG00000090104	RGS1	6.69	7.55e-25	0.000
1	ENSG00000130755	GMFG	6.39	3.51e-26	0.000
1	ENSG00000132464	ENAM	5.91	6.01e-36	0.000
1	ENSG00000211598	IGKV4-1	7.55	6.08e-24	0.002
1	ENSG00000239855	IGKV1-6	8.35	7.95e-24	0.003
1	ENSG00000108405	P2RX1	6.87	6.97e-24	0.003
1	ENSG00000241755	IGKV1-9	7.32	1.45e-23	0.006
1	ENSG00000143297	FCRL5	7.69	3.93e-23	0.01
1	ENSG00000251546	IGKV1D-39	7.37	3.64e-23	0.01
1	ENSG00000211895	IGHA1	8.42	1.08e-22	0.016
1	ENSG00000077943	ITGA8	6.94	2.64e-22	0.024
1	ENSG00000242371	IGKV1-39	6.96	3.27e-22	0.025
1	ENSG00000253701	ENSG00000253701	6.81	5.21e-22	0.031
1	ENSG00000211947	IGHV3-21	5.5	7.46e-22	0.033
1	ENSG00000268027	ENSG00000268027	6.44	9.99e-22	0.035

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000169413	RNASE6	6.81	1.1e-21	0.036
1	ENSG00000243063	IGKV3-7	6.28	1.01e-21	0.036
1	ENSG00000254709	IGLL5	7.77	1.23e-21	0.04
1	ENSG00000170421	KRT8	-7.49	3.98e-27	0.000
2	ENSG00000167900	TK1	6.63	3.11e-36	0.000
2	ENSG00000168496	FEN1	6.62	2.44e-34	0.000
2	ENSG00000143476	DTL	6.6	1.29e-34	0.000
2	ENSG00000100162	CENPM	6.55	7.58e-37	0.000
2	ENSG00000184357	HIST1H1B	6.38	3.86e-22	0.000
2	ENSG00000137804	NUSAP1	6.34	4.26e-37	0.000
2	ENSG00000170312	CDK1	6.25	9.84e-31	0.000
2	ENSG00000171848	RRM2	6	3.16e-20	0.000
2	ENSG00000012048	BRCA1	5.97	8.34e-43	0.000
2	ENSG00000189057	FAM111B	5.91	8.28e-33	0.000
2	ENSG00000094804	CDC6	5.85	1.37e-26	0.000
2	ENSG00000105011	ASF1B	5.83	4.5e-27	0.000
2	ENSG00000175063	UBE2C	5.78	4.9e-26	0.000
2	ENSG00000111247	RAD51AP1	5.67	3.1e-30	0.000
2	ENSG00000198554	WDHD1	5.65	1.07e-21	0.000
2	ENSG00000163808	KIF15	5.52	5.31e-23	0.000
2	ENSG00000166803	KIAA0101	5.49	2.62e-35	0.000
2	ENSG00000092853	CLSPN	5.43	3.74e-31	0.000
2	ENSG00000065328	MCM10	5.31	2.89e-21	0.000
2	ENSG00000156970	BUB1B	5.29	1.73e-20	0.000
2	ENSG00000151725	CENPU	5.19	6.05e-24	0.000
2	ENSG00000119969	HELLS	5.09	7.42e-20	0.000
2	ENSG00000176208	ATAD5	4.76	9.1e-20	0.000
2	ENSG00000121152	NCAPH	6.09	8e-18	0.003
2	ENSG00000011426	ANLN	5.45	1.14e-18	0.003
2	ENSG00000163918	RFC4	5.18	1.42e-18	0.003
2	ENSG00000137310	TCF19	5.07	3.53e-19	0.003
2	ENSG00000196584	XRCC2	5.01	1.62e-18	0.003
2	ENSG00000121211	MND1	4.62	3.08e-19	0.003
2	ENSG00000142731	PLK4	4.59	7.55e-18	0.003
2	ENSG00000140525	FANCI	5.04	1.75e-17	0.005
2	ENSG00000158402	CDC25C	4.74	4.36e-17	0.005
2	ENSG00000051180	RAD51	5.25	1.14e-16	0.008
2	ENSG00000167513	CDT1	4.92	3.51e-16	0.012
2	ENSG00000176890	TYMS	4.79	2.74e-16	0.012
2	ENSG00000143228	NUF2	4.43	2.66e-16	0.012
2	ENSG00000093009	CDC45	5.97	5.98e-16	0.02
2	ENSG00000152253	SPC25	4.67	5.77e-16	0.02

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
2	ENSG00000160957	RECQL4	4.41	5.68e-16	0.02
2	ENSG00000183856	IQGAP3	4.66	1.18e-15	0.027
2	ENSG00000101412	E2F1	5.06	1.51e-15	0.031
2	ENSG00000131153	GINS2	5	2.56e-15	0.036
2	ENSG00000136982	DSCC1	4.52	4.16e-15	0.038
3	ENSG00000142583	SLC2A5	-4.15	2.25e-15	0.000
3	ENSG00000160870	CYP3A7	-4.64	5.63e-15	0.000
3	ENSG00000160868	CYP3A4	-5.14	2.5e-15	0.000
3	ENSG00000172689	MS4A10	-5.44	9.5e-15	0.000
3	ENSG00000165272	AQP3	-5.6	6.38e-20	0.000
3	ENSG00000084674	APOB	-5.81	1.57e-17	0.000
3	ENSG00000076351	SLC46A1	-5.87	7.06e-18	0.000
3	ENSG00000156222	SLC28A1	-5.89	9.2e-16	0.000
3	ENSG00000173237	C11orf86	-6.57	1.18e-18	0.000
3	ENSG00000118137	APOA1	-6.96	4.43e-26	0.000
3	ENSG00000187288	CIDEC	-7.01	1.69e-20	0.000
3	ENSG00000198788	MUC2	-7.06	1.44e-18	0.000
3	ENSG00000110245	APOC3	-7.27	2.48e-19	0.000
3	ENSG00000016490	CLCA1	-7.45	2.92e-15	0.000
3	ENSG00000110244	APOA4	-7.61	3.16e-23	0.000
3	ENSG00000169903	TM4SF4	-5.12	1.56e-13	0.001
3	ENSG00000184163	FAM132A	-5.25	2.16e-13	0.001
3	ENSG00000163295	ALPI	-5.61	2.31e-13	0.001
3	ENSG00000143595	AQP10	-6.33	1.6e-13	0.001
3	ENSG00000104808	DHDH	-6.14	3.61e-13	0.002
3	ENSG00000132465	JCHAIN	-8.07	5.05e-13	0.002
3	ENSG00000145384	FABP2	-2.14	1.89e-12	0.004
3	ENSG00000115850	LCT	-4.12	2.71e-12	0.005
3	ENSG00000179914	ITLN1	-6.2	3.11e-12	0.005
3	ENSG00000160180	TFF3	-3.46	4.5e-12	0.008
3	ENSG00000130055	GDPD2	-5.49	7.42e-12	0.008
3	ENSG00000228146	CASP16P	-5.71	1.95e-11	0.016
3	ENSG00000184709	LRRC26	-5.24	2.13e-11	0.017
3	ENSG00000182156	ENPP7	-4.99	2.96e-11	0.019
3	ENSG00000174992	ZG16	-4.01	4.4e-11	0.022
3	ENSG00000229391	HLA-DRB6	-2.6	5.32e-11	0.024
3	ENSG00000100162	CENPM	-5.18	8.88e-11	0.026
3	ENSG00000100170	SLC5A1	-2.72	3.1e-10	0.038
3	ENSG00000094804	CDC6	-4.76	4.19e-10	0.045
3	ENSG00000141434	MEP1B	-4.55	4.27e-10	0.046
4	ENSG00000110244	APOA4	8.77	1.97e-76	0.000
4	ENSG00000110245	APOC3	8.49	8.91e-72	0.000

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000118137	APOA1	8.12	1.6e-93	0.000
4	ENSG00000187288	CIDEC	7.99	1.23e-89	0.000
4	ENSG00000173237	C11orf86	7.35	1.68e-64	0.000
4	ENSG00000143595	AQP10	7.17	1.33e-35	0.000
4	ENSG00000084674	APOB	6.99	1.23e-47	0.000
4	ENSG00000163295	ALPI	6.95	8.67e-39	0.000
4	ENSG00000172689	MS4A10	6.94	2.74e-61	0.000
4	ENSG00000182156	ENPP7	6.36	2.18e-37	0.000
4	ENSG00000160868	CYP3A4	6.31	8.88e-40	0.000
4	ENSG00000130055	GDPD2	6.23	1.57e-28	0.000
4	ENSG00000076351	SLC46A1	6.1	7.84e-40	0.000
4	ENSG00000156222	SLC28A1	6.02	6.86e-31	0.000
4	ENSG00000023839	ABCC2	5.91	4.66e-26	0.000
4	ENSG00000169906	S100G	5.89	2.84e-26	0.000
4	ENSG00000169903	TM4SF4	5.87	1.99e-27	0.000
4	ENSG00000173406	DAB1	5.59	6e-27	0.000
4	ENSG00000160870	CYP3A7	5.57	8.34e-46	0.000
4	ENSG00000184163	FAM132A	5.47	3.94e-24	0.000
4	ENSG00000115850	LCT	5.41	3.96e-32	0.000
4	ENSG00000188611	ASAH2	5.41	7.07e-24	0.000
4	ENSG00000111701	APOBEC1	5.36	3.03e-26	0.000
4	ENSG00000142583	SLC2A5	4.72	2.87e-29	0.000
4	ENSG00000159640	ACE	4.66	1.11e-23	0.000
4	ENSG00000213822	CEACAM18	4.64	2.28e-30	0.000
4	ENSG00000136872	ALDOB	4.1	4.98e-27	0.000
4	ENSG00000100170	SLC5A1	3.98	1.44e-28	0.000
4	ENSG00000174028	FAM3C2	3.67	1.58e-25	0.000
4	ENSG00000242366	UGT1A8	3.47	7.28e-25	0.000
4	ENSG00000227471	AKR1B15	3.41	1.37e-26	0.000
4	ENSG00000229391	HLA-DRB6	3.17	1.21e-27	0.000
4	ENSG00000145384	FABP2	2.75	8.03e-24	0.000
4	ENSG00000197142	ACSL5	2.58	9.97e-27	0.000
4	ENSG00000214782	MS4A18	6.92	5.17e-21	0.001
4	ENSG00000256612	CYP2B7P	6.13	1.01e-21	0.001
4	ENSG00000176919	C8G	5.94	3.88e-20	0.001
4	ENSG00000141434	MEP1B	5.41	2.05e-20	0.001
4	ENSG00000257335	MGAM	5.19	4.83e-21	0.001
4	ENSG00000138075	ABCG5	5.15	4.55e-20	0.001
4	ENSG00000229005	HNF4A-AS1	4.9	5.15e-23	0.001
4	ENSG00000088386	SLC15A1	4.68	8.11e-20	0.001
4	ENSG00000107798	LIPA	4.62	9.78e-21	0.001
4	ENSG00000091138	SLC26A3	4.54	1.8e-19	0.001

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000250298	GIMD1	3.8	2.43e-19	0.001
4	ENSG00000138823	MTTP	3.45	6.59e-19	0.001
4	ENSG00000169894	MUC3A	3.28	1.36e-20	0.001
4	ENSG00000173702	MUC13	3.26	3.14e-19	0.001
4	ENSG00000162482	AKR7A3	3.05	2.25e-19	0.001
4	ENSG00000198074	AKR1B10	2.99	4.27e-20	0.001
4	ENSG00000114786	ABHD14A-ACY1	2.53	5.49e-20	0.001
4	ENSG00000165140	FBP1	2.4	6.11e-20	0.001
4	ENSG00000165828	PRAP1	2.22	7.22e-23	0.001
4	ENSG00000095932	SMIM24	2.18	8.74e-19	0.001
4	ENSG00000182938	OTOP3	5.02	2.23e-18	0.002
4	ENSG00000074276	CDHR2	4.66	4.66e-18	0.002
4	ENSG00000133328	HRASLS2	4.57	1.1e-17	0.002
4	ENSG00000104267	CA2	4.03	1.32e-18	0.002
4	ENSG00000104808	DHDH	5.6	2.93e-17	0.003
4	ENSG00000108242	CYP2C18	3.14	6.83e-17	0.005
4	ENSG00000197241	SLC2A7	5.21	2.83e-16	0.007
4	ENSG00000100197	CYP2D6	5.18	1.95e-16	0.007
4	ENSG00000154646	TMPRSS15	4.57	2.42e-16	0.007
4	ENSG00000204978	ERICH4	4.33	1.87e-16	0.007
4	ENSG00000250722	SEPP1	4.04	1.05e-16	0.007
4	ENSG00000124253	PCK1	4.04	2.37e-16	0.007
4	ENSG00000124299	PEPD	3.29	2.51e-16	0.007
4	ENSG00000115310	RTN4	2.89	2.88e-16	0.007
4	ENSG00000147804	SLC39A4	2.69	2.52e-16	0.007
4	ENSG00000254521	SIGLEC12	3.65	4.72e-16	0.009
4	ENSG00000019582	CD74	1.74	5.68e-16	0.009
4	ENSG00000198668	CALM1	1.55	4.66e-16	0.009
4	ENSG00000131482	G6PC	4.62	7.03e-16	0.01
4	ENSG00000099834	CDHR5	2.85	7.07e-16	0.01
4	ENSG00000120057	SFRP5	4.61	7.55e-16	0.011
4	ENSG00000118777	ABCG2	4.51	8.5e-16	0.012
4	ENSG00000174827	PDZK1	3.79	1.48e-15	0.012
4	ENSG00000104537	ANXA13	3.32	1.49e-15	0.012
4	ENSG00000130234	ACE2	4.23	2.18e-15	0.013
4	ENSG00000008517	IL32	3.75	2.01e-15	0.013
4	ENSG00000119125	GDA	3.31	1.73e-15	0.013
4	ENSG00000196502	SULT1A1	2.66	2.22e-15	0.013
4	ENSG00000090402	SI	2.29	1.61e-15	0.013
4	ENSG00000100889	PCK2	3.12	3.03e-15	0.014
4	ENSG00000166959	MS4A8	3.2	3.41e-15	0.015
4	ENSG00000153233	PTPRR	5.32	4.04e-15	0.016

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000198814	GK	3.7	3.65e-15	0.016
4	ENSG00000214110	LDHAP4	2.49	4.36e-15	0.017
4	ENSG00000161533	ACOX1	2.65	4.85e-15	0.018
4	ENSG00000118515	SGK1	4.56	5.69e-15	0.019
4	ENSG00000185000	DGAT1	2.97	5.51e-15	0.019
4	ENSG00000114455	HHLA2	4.63	8.71e-15	0.02
4	ENSG00000118094	TREH	4.47	1.26e-14	0.02
4	ENSG00000256162	SMLR1	4.07	2.36e-14	0.02
4	ENSG00000173221	GLRX	3.73	1.68e-14	0.02
4	ENSG00000174358	SLC6A19	3.62	1.4e-14	0.02
4	ENSG00000196839	ADA	3.58	2.14e-14	0.02
4	ENSG00000266714	MYO15B	3.39	1.55e-14	0.02
4	ENSG00000198502	HLA-DRB5	3.07	1.25e-14	0.02
4	ENSG00000196937	FAM3C	3.05	2.07e-14	0.02
4	ENSG00000166866	MYO1A	2.73	1.06e-14	0.02
4	ENSG00000243989	ACY1	2.38	2.41e-14	0.02
4	ENSG00000136156	ITM2B	1.23	3.28e-14	0.021
4	ENSG00000007216	SLC13A2	3.43	5.38e-14	0.022
4	ENSG00000211454	AKR7L	3	4.13e-14	0.022
4	ENSG00000134333	LDHA	2.21	4.41e-14	0.022
4	ENSG00000140465	CYP1A1	6.1	1.33e-13	0.023
4	ENSG00000178462	TUBAL3	4.69	8.27e-14	0.023
4	ENSG00000186115	CYP4F2	4.19	1.11e-13	0.023
4	ENSG00000242574	HLA-DMB	4.11	1.54e-13	0.023
4	ENSG00000163959	SLC51A	4.08	1.63e-13	0.023
4	ENSG00000168955	TM4SF20	3.69	1.5e-13	0.023
4	ENSG00000196126	HLA-DRB1	2.54	1.22e-13	0.023
4	ENSG00000159228	CBR1	2.25	7.12e-14	0.023
4	ENSG00000107223	EDF1	1.36	1.56e-13	0.023
4	ENSG00000167701	GPT	5.04	2e-13	0.025
4	ENSG00000244503	ENSG00000244503	5.47	2.1e-13	0.026
4	ENSG00000142484	TM4SF5	2.36	2.09e-13	0.026
4	ENSG00000126218	F10	5.34	2.12e-13	0.027
4	ENSG00000213918	DNASE1	2.59	2.23e-13	0.027
4	ENSG00000167468	GPX4	2.12	2.27e-13	0.027
4	ENSG00000228146	CASP16P	5.54	3.3e-13	0.033
4	ENSG00000160179	ABCG1	4.85	3.41e-13	0.033
4	ENSG00000131910	NR0B2	3.83	3.42e-13	0.033
4	ENSG00000176273	SLC35G1	3.35	3.61e-13	0.034
4	ENSG00000109099	PMP22	4.48	3.69e-13	0.035
4	ENSG00000204287	HLA-DRA	2.44	3.66e-13	0.035
4	ENSG00000117984	CTSD	2.26	3.88e-13	0.036

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000155465	SLC7A7	4.12	4.44e-13	0.038
4	ENSG00000223865	HLA-DPB1	4.69	6.65e-13	0.048
4	ENSG00000234906	APOC2	2.61	6.74e-13	0.049
4	ENSG00000102837	OLFM4	-7.27	4.71e-31	0.000
4	ENSG00000064651	SLC12A2	-4.35	6.1e-22	0.001
4	ENSG00000213853	EMP2	-6.73	1.93e-20	0.001
4	ENSG00000144354	CDCA7	-6.94	2.82e-19	0.001
4	ENSG00000172016	REG3A	-7.62	1.27e-20	0.001
4	ENSG00000197756	RPL37A	-1.34	1.1e-17	0.002
4	ENSG00000071082	RPL31	-1.47	9.32e-18	0.002
4	ENSG00000149273	RPS3	-1.67	1.88e-18	0.002
4	ENSG00000140988	RPS2	-1.72	4.31e-17	0.003
4	ENSG00000167526	RPL13	-1.29	1.86e-16	0.007
4	ENSG00000105640	RPL18A	-1.4	2e-16	0.007
4	ENSG00000240342	RPS2P5	-2.15	3.32e-16	0.007
4	ENSG00000187514	PTMA	-1.43	1.53e-15	0.012
4	ENSG00000142937	RPS8	-1.6	1.46e-15	0.012
4	ENSG00000232573	RPL3P4	-1.59	2.17e-15	0.013
4	ENSG00000137154	RPS6	-1.4	3.03e-15	0.014
4	ENSG00000172023	REG1B	-3.11	3.03e-15	0.014
4	ENSG00000112306	RPS12	-1.52	7.22e-15	0.02
4	ENSG00000242299	ENSG00000242299	-1.65	1.33e-14	0.02
4	ENSG00000096384	HSP90AB1	-2.43	1.01e-14	0.02
4	ENSG00000107159	CA9	-6.56	6.38e-15	0.02
4	ENSG00000174444	RPL4	-1.91	3.95e-14	0.022
4	ENSG00000229638	RPL4P4	-1.99	6.06e-14	0.022
4	ENSG00000109475	RPL34	-1.16	1.22e-13	0.023
4	ENSG00000231500	RPS18	-1.25	1.62e-13	0.023
4	ENSG00000116299	KIAA1324	-3.8	8.95e-14	0.023
4	ENSG00000148303	RPL7A	-1.45	1.95e-13	0.025
4	ENSG00000198242	RPL23A	-0.989	3.07e-13	0.032
4	ENSG00000125691	RPL23	-1.45	3.91e-13	0.036
4	ENSG00000232176	ENSG00000232176	-1.77	5.84e-13	0.043
4	ENSG00000143947	RPS27A	-1.28	6.37e-13	0.044
4	ENSG00000130255	RPL36	-1.1	6.68e-13	0.049
5	ENSG00000016490	CLCA1	10.3	1.27e-79	0.000
5	ENSG00000198788	MUC2	10	1.63e-139	0.000
5	ENSG00000179914	ITLN1	9.25	7.72e-67	0.000
5	ENSG00000162998	FRZB	9	1.51e-70	0.000
5	ENSG00000158764	ITLN2	8.74	2.12e-43	0.000
5	ENSG00000064787	BCAS1	8.05	1e-38	0.000
5	ENSG00000197520	FAM177B	7.33	7.96e-47	0.000

Table A.11 Duodenal differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
5	ENSG00000122711	SPINK4	7.29	3.32e-36	0.000
5	ENSG00000135773	CAPN9	7.23	4.29e-34	0.000
5	ENSG00000250591	PRSS3P1	7.19	6.9e-27	0.000
5	ENSG00000181104	F2R	7.18	5.24e-27	0.000
5	ENSG00000130035	GALNT8	7.09	2.03e-31	0.000
5	ENSG00000074527	NTN4	6.98	4.05e-26	0.000
5	ENSG00000106178	CCL24	6.97	1.32e-28	0.000
5	ENSG00000174236	REP15	6.61	1e-35	0.000
5	ENSG00000160180	TFF3	6.5	1.39e-66	0.000
5	ENSG00000134193	REG4	6.12	1.11e-36	0.000
5	ENSG00000198691	ABCA4	7.59	1.59e-22	0.002
5	ENSG00000172238	ATOH1	6.6	1.59e-22	0.002
5	ENSG00000259974	LINC00261	6.59	1.97e-23	0.002
5	ENSG00000102003	SYP	5.74	2.62e-23	0.002
5	ENSG00000163515	RETNLB	8.27	1.08e-21	0.004
5	ENSG00000204983	PRSS1	7.03	2.82e-21	0.004
5	ENSG00000184709	LRRC26	6.4	1.63e-21	0.004
5	ENSG00000136542	GALNT5	6.57	9.75e-21	0.009
5	ENSG00000129538	RNASE1	5.84	7.75e-20	0.018
5	ENSG00000138119	MYOF	5.79	2.31e-19	0.025
5	ENSG00000214814	FER1L6	6.68	3e-19	0.028

Table A.11: Duodenal differentially expressed genes. The table displays the genes identified as significant (FDR <0.05) in differential expression analysis of the duodenal single cell clusters. Genes are ordered by cluster and significance and the log2 fold-change in the cluster compared to the rest of the duodenal single cell libraries is listed for each gene. The average expression across across all considered single cell libraries in log2 counts per million is also listed.

A.2.6 Differentially expressed genes across Barrett's tissue clusters

Table A.12 Barrett's esophagus differentially expressed genes.

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000179914	ITLN1	6.73	2.65e-50	0.000
1	ENSG00000174562	KLK15	6.36	3.1e-35	0.000
1	ENSG00000174992	ZG16	6.33	4.54e-36	0.000
1	ENSG00000158764	ITLN2	6.28	3.55e-48	0.000
1	ENSG00000162998	FRZB	6.26	5.7e-30	0.000
1	ENSG00000116176	TPSG1	6.18	1.17e-41	0.000
1	ENSG00000130035	GALNT8	6.12	1.86e-45	0.000
1	ENSG00000198788	MUC2	5.64	2.83e-54	0.000
1	ENSG00000100302	RASD2	5.44	1.41e-30	0.000
1	ENSG00000198719	DLL1	5.13	2.94e-26	0.000
1	ENSG00000167748	KLK1	4.85	1.42e-36	0.000
1	ENSG00000122711	SPINK4	4.84	1.92e-40	0.000
1	ENSG00000142515	KLK3	4.8	3.55e-54	0.000
1	ENSG00000144485	HES6	4.69	1.41e-28	0.000
1	ENSG00000184709	LRRC26	4.52	3.02e-34	0.000
1	ENSG00000174236	REP15	4.48	3.45e-38	0.000
1	ENSG00000076864	RAP1GAP	4.41	2.85e-30	0.000
1	ENSG00000070526	ST6GALNAC1	3.41	9.26e-54	0.000
1	ENSG00000160180	TFF3	1.8	1.39e-29	0.000
1	ENSG00000142669	SH3BGRL3	1.52	1.56e-26	0.000
1	ENSG00000031081	ARHGAP31	4.62	4.76e-24	0.001
1	ENSG00000077232	DNAJC10	2.2	2.6e-24	0.001
1	ENSG00000169442	CD52	5.01	9.47e-24	0.002
1	ENSG00000148180	GSN	2.66	2.06e-21	0.036
1	ENSG00000120306	CYSTM1	-2.78	7.48e-29	0.000
1	ENSG00000135052	GOLM1	-3.82	4.49e-26	0.000
1	ENSG00000104267	CA2	-6.3	2.05e-25	0.000
1	ENSG00000248144	ADH1C	-7.54	9.7e-37	0.000
1	ENSG00000102837	OLFM4	-7.76	1.01e-25	0.000
1	ENSG00000001626	CFTR	-6.37	1.96e-24	0.001
1	ENSG00000112297	AIM1	-4.22	7.98e-22	0.024
1	ENSG00000104783	KCNN4	-6.81	2.07e-21	0.036
2	ENSG00000100604	CHGA	8.81	1.73e-70	0.000
2	ENSG00000109472	CPE	8.43	3.66e-85	0.000
2	ENSG00000118271	TTR	8.32	4.28e-74	0.000
2	ENSG00000116774	OLFML3	8.06	5.33e-48	0.000
2	ENSG00000096395	MLN	8.04	3.43e-41	0.000
2	ENSG00000133636	NTS	7.8	2.02e-33	0.000

Table A.12 Barrett's differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
2	ENSG00000089199	CHGB	7.8	9.68e-92	0.000
2	ENSG00000079689	SCGN	7.39	2.99e-60	0.000
2	ENSG00000104112	SCG3	7.39	2.66e-76	0.000
2	ENSG00000163499	CRYBA2	7.1	1.08e-43	0.000
2	ENSG00000131711	MAP1B	6.78	4.2e-58	0.000
2	ENSG00000145362	ANK2	6.62	2.39e-39	0.000
2	ENSG00000117152	RGS4	6.59	4.68e-42	0.000
2	ENSG00000102466	FGF14	6.5	4.43e-33	0.000
2	ENSG00000132639	SNAP25	6.33	1.39e-37	0.000
2	ENSG00000147231	CXorf57	6.3	1.79e-35	0.000
2	ENSG00000159409	CELF3	6.27	3.64e-57	0.000
2	ENSG00000168280	KIF5C	6.15	1.3e-44	0.000
2	ENSG00000182901	RGS7	5.37	3.84e-40	0.000
2	ENSG00000152932	RAB3C	5.32	6.19e-38	0.000
2	ENSG00000176840	MIR7-3HG	5.25	1.05e-33	0.000
2	ENSG00000128536	CDHR3	5.84	7.99e-33	0.001
2	ENSG00000179921	GPBAR1	6.21	1.14e-31	0.005
2	ENSG00000171951	SCG2	7.36	3.54e-31	0.006
2	ENSG00000171759	PAH	7.29	2.17e-31	0.006
2	ENSG00000102109	PCSK1N	7.08	1.99e-31	0.006
2	ENSG00000129167	TPH1	7.57	4.35e-31	0.007
2	ENSG00000197584	KCNMB2	5.31	9.4e-31	0.008
2	ENSG00000153002	CPB1	6.94	2.37e-30	0.011
2	ENSG00000225792	ENSG00000225792	4.99	2.88e-30	0.011
2	ENSG00000136999	NOV	5.62	3.15e-29	0.016
2	ENSG00000115263	GCG	7.78	3.36e-28	0.033
2	ENSG00000134198	TSPAN2	5.8	3.12e-28	0.033
2	ENSG00000163497	FEV	6.52	1.77e-27	0.048
4	ENSG00000196616	ADH1B	2.18	1.05e-15	0.01
4	ENSG00000248144	ADH1C	2.88	3.46e-15	0.024
4	ENSG00000269028	MTRNR2L12	0.888	1.53e-14	0.04
4	ENSG00000198363	ASPH	-1.92	2.38e-19	0.000
4	ENSG00000110492	MDK	-2.31	5.08e-22	0.000
4	ENSG00000070526	ST6GALNAC1	-2.54	1.2e-25	0.000
4	ENSG00000167552	TUBA1A	-3.42	1.89e-34	0.000
4	ENSG00000259974	LINC00261	-3.44	2.88e-20	0.000
4	ENSG00000142515	KLK3	-3.7	6.27e-22	0.000
4	ENSG00000139998	RAB15	-3.88	6.83e-19	0.000
4	ENSG00000184709	LRRC26	-3.88	1.07e-20	0.000
4	ENSG00000143001	TMEM61	-4.18	2.57e-22	0.000
4	ENSG00000118193	KIF14	-4.45	4.39e-21	0.000
4	ENSG00000198788	MUC2	-4.5	7.66e-28	0.000

Table A.12 Barrett's differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000122711	SPINK4	-4.57	9.25e-30	0.000
4	ENSG00000161888	SPC24	-4.62	1.04e-18	0.000
4	ENSG00000112984	KIF20A	-4.7	6.58e-21	0.000
4	ENSG00000167748	KLK1	-4.76	1.19e-28	0.000
4	ENSG00000136854	STXBP1	-4.78	4.82e-27	0.000
4	ENSG00000090889	KIF4A	-5.01	6.94e-21	0.000
4	ENSG00000137807	KIF23	-5.08	4.16e-21	0.000
4	ENSG00000144485	HES6	-5.08	2.65e-29	0.000
4	ENSG00000171303	KCNK3	-5.11	7.82e-19	0.000
4	ENSG00000080986	NDC80	-5.11	1.53e-22	0.000
4	ENSG00000134057	CCNB1	-5.19	1.03e-19	0.000
4	ENSG00000171848	RRM2	-5.2	6.17e-19	0.000
4	ENSG00000076382	SPAG5	-5.22	2.75e-22	0.000
4	ENSG00000142945	KIF2C	-5.28	1.08e-21	0.000
4	ENSG00000107954	NEURL1	-5.32	6.34e-36	0.000
4	ENSG00000188175	HEPACAM2	-5.61	1.39e-21	0.000
4	ENSG00000158764	ITLN2	-5.61	1.73e-24	0.000
4	ENSG00000089199	CHGB	-5.69	7.94e-19	0.000
4	ENSG00000130035	GALNT8	-5.71	1.41e-25	0.000
4	ENSG00000116176	TPSG1	-5.81	1.84e-25	0.000
4	ENSG00000100302	RASD2	-5.87	1.02e-28	0.000
4	ENSG00000036565	SLC18A1	-5.96	4.52e-20	0.000
4	ENSG00000185002	RFX6	-6.02	1.8e-21	0.000
4	ENSG00000066279	ASPM	-6.04	6.26e-25	0.000
4	ENSG00000174992	ZG16	-6.22	3.79e-25	0.000
4	ENSG00000108551	RASD1	-6.32	9.64e-40	0.000
4	ENSG00000198719	DLL1	-6.69	1.49e-39	0.000
4	ENSG00000186474	KLK12	-6.72	3.51e-37	0.000
4	ENSG00000109472	CPE	-6.74	2.01e-22	0.000
4	ENSG00000179914	ITLN1	-6.96	1.58e-38	0.000
4	ENSG00000016490	CLCA1	-7.21	2.72e-22	0.000
4	ENSG00000100604	CHGA	-7.85	2.83e-31	0.000
4	ENSG00000124766	SOX4	-2.19	2.39e-17	0.001
4	ENSG00000076864	RAP1GAP	-3.65	1.56e-17	0.001
4	ENSG00000152253	SPC25	-3.99	2.41e-18	0.001
4	ENSG00000101057	MYBL2	-4.09	1.69e-17	0.001
4	ENSG00000183856	IQGAP3	-4.28	5.45e-18	0.001
4	ENSG00000169679	BUB1	-4.62	5.64e-18	0.001
4	ENSG00000137804	NUSAP1	-4.84	3.14e-18	0.001
4	ENSG00000131747	TOP2A	-5.24	2.16e-17	0.001
4	ENSG00000117399	CDC20	-5.38	2.38e-18	0.001
4	ENSG00000167964	RAB26	-4.1	8.98e-17	0.002

Table A.12 Barrett's differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
4	ENSG00000117650	NEK2	-4.16	2.64e-16	0.002
4	ENSG00000185905	C16orf54	-4.36	1.06e-16	0.002
4	ENSG00000255474	ENSG00000255474	-4.41	1.08e-16	0.002
4	ENSG00000126787	DLGAP5	-4.44	2.54e-17	0.002
4	ENSG00000169607	CKAP2L	-4.45	4.35e-17	0.002
4	ENSG00000145362	ANK2	-5.05	1.08e-16	0.002
4	ENSG00000175426	PCSK1	-5.36	4.43e-17	0.002
4	ENSG00000006071	ABCC8	-5.42	9.41e-17	0.002
4	ENSG00000163497	FEV	-5.45	7.14e-17	0.002
4	ENSG00000118271	TTR	-5.9	2.3e-16	0.002
4	ENSG00000102109	PCSK1N	-6.15	3.73e-17	0.002
4	ENSG00000072571	HMMR	-4.05	2.96e-16	0.003
4	ENSG00000174562	KLK15	-5.24	3.25e-16	0.004
4	ENSG00000079689	SCGN	-5.38	4.01e-16	0.005
4	ENSG00000112379	ARFGEF3	-1.89	4.85e-16	0.007
4	ENSG00000157456	CCNB2	-4.4	5.36e-16	0.007
4	ENSG00000156970	BUB1B	-4.4	5.01e-16	0.007
4	ENSG00000168078	PBK	-4.48	5.71e-16	0.007
4	ENSG00000134954	ETS1	-3.66	7.98e-16	0.009
4	ENSG00000170312	CDK1	-5.2	8.62e-16	0.009
4	ENSG00000115163	CENPA	-3.73	9.56e-16	0.01
4	ENSG00000145198	VWA5B2	-5.17	1.04e-15	0.01
4	ENSG00000163808	KIF15	-4.39	1.14e-15	0.012
4	ENSG00000088325	TPX2	-4.57	1.3e-15	0.014
4	ENSG00000122859	NEUROG3	-5.95	1.99e-15	0.017
4	ENSG00000115756	HPCAL1	-2.38	4.45e-15	0.025
4	ENSG00000162998	FRZB	-5.18	3.61e-15	0.025
4	ENSG00000129195	FAM64A	-4.46	5.89e-15	0.027
4	ENSG00000102003	SYP	-3.46	7.99e-15	0.03
4	ENSG00000163618	CADPS	-3.48	8.3e-15	0.03
4	ENSG00000147488	ST18	-4.8	7.9e-15	0.03
4	ENSG00000126803	HSPA2	-3.65	1.03e-14	0.034
4	ENSG00000178999	AURKB	-4.43	1.07e-14	0.034
4	ENSG00000143228	NUF2	-3.92	1.44e-14	0.04
4	ENSG00000089685	BIRC5	-4.03	1.65e-14	0.041
4	ENSG00000131711	MAP1B	-4.73	1.79e-14	0.043
4	ENSG00000011347	SYT7	-3.97	1.92e-14	0.046
5	ENSG00000066279	ASPM	7.87	9.55e-178	0.000
5	ENSG00000131747	TOP2A	7.32	6.05e-67	0.000
5	ENSG00000171848	RRM2	7.28	1.39e-77	0.000
5	ENSG00000117399	CDC20	7.28	1.43e-71	0.000
5	ENSG00000170312	CDK1	7.11	2.06e-64	0.000

Table A.12 Barrett's differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
5	ENSG00000134057	CCNB1	6.81	7.51e-63	0.000
5	ENSG00000137804	NUSAP1	6.74	1.5e-71	0.000
5	ENSG00000088325	TPX2	6.72	1.7e-70	0.000
5	ENSG00000169679	BUB1	6.62	1.07e-107	0.000
5	ENSG00000137807	KIF23	6.55	7.78e-79	0.000
5	ENSG00000178999	AURKB	6.5	2.19e-57	0.000
5	ENSG00000168078	PBK	6.42	1.05e-100	0.000
5	ENSG00000175063	UBE2C	6.42	7.9e-39	0.000
5	ENSG00000157456	CCNB2	6.35	2.05e-62	0.000
5	ENSG00000156970	BUB1B	6.35	1.36e-77	0.000
5	ENSG00000112984	KIF20A	6.31	2.43e-81	0.000
5	ENSG00000163808	KIF15	6.26	1.92e-69	0.000
5	ENSG00000090889	KIF4A	6.26	1.34e-65	0.000
5	ENSG00000126787	DLGAP5	6.25	6.57e-79	0.000
5	ENSG00000024526	DEPDC1	6.23	1.14e-68	0.000
5	ENSG00000089685	BIRC5	6.21	2.66e-69	0.000
5	ENSG00000121152	NCAPH	6.19	1.29e-59	0.000
5	ENSG00000142945	KIF2C	6.18	2.23e-42	0.000
5	ENSG00000148773	MKI67	6.15	3.51e-35	0.000
5	ENSG00000117724	CENPF	6.14	2.69e-37	0.000
5	ENSG00000161888	SPC24	6.12	9.89e-62	0.000
5	ENSG00000080986	NDC80	6.09	5.22e-47	0.000
5	ENSG00000118193	KIF14	6.08	2.17e-77	0.000
5	ENSG00000169607	CKAP2L	6.05	1.17e-56	0.000
5	ENSG00000072571	HMMR	6	8.76e-89	0.000
5	ENSG00000129195	FAM64A	5.99	1.82e-39	0.000
5	ENSG00000123485	HJURP	5.99	5.06e-51	0.000
5	ENSG00000184357	HIST1H1B	5.93	4.74e-32	0.000
5	ENSG00000166851	PLK1	5.91	5.43e-36	0.000
5	ENSG00000117650	NEK2	5.87	4.81e-67	0.000
5	ENSG00000165304	MELK	5.83	6.31e-47	0.000
5	ENSG00000101057	MYBL2	5.8	2.72e-71	0.000
5	ENSG00000145386	CCNA2	5.78	1.07e-45	0.000
5	ENSG00000138778	CENPE	5.78	4e-40	0.000
5	ENSG00000186185	KIF18B	5.77	5.01e-54	0.000
5	ENSG00000143228	NUF2	5.76	1.11e-61	0.000
5	ENSG00000171241	SHCBP1	5.72	6.36e-46	0.000
5	ENSG00000183856	IQGAP3	5.7	3.03e-49	0.000
5	ENSG00000109805	NCAPG	5.69	4.66e-57	0.000
5	ENSG00000138180	CEP55	5.68	5.98e-41	0.000
5	ENSG00000075218	GTSE1	5.67	4.55e-43	0.000
5	ENSG00000152253	SPC25	5.62	7.18e-68	0.000

Table A.12 Barrett's differentially expressed genes, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
5	ENSG00000100526	CDKN3	5.57	1.34e-37	0.000
5	ENSG00000115163	CENPA	5.5	1.3e-64	0.000
5	ENSG00000129810	SGOL1	5.38	7.09e-37	0.000
5	ENSG00000154839	SKA1	5.37	8.79e-38	0.000
5	ENSG00000109674	NEIL3	5.33	8.59e-38	0.000
5	ENSG00000163535	SGOL2	5.32	1.11e-33	0.000
5	ENSG00000068489	PRR11	5.24	5.66e-39	0.000
5	ENSG00000198901	PRC1	5.23	3.91e-31	0.000
5	ENSG00000138160	KIF11	5.17	5.26e-31	0.000
5	ENSG00000164109	MAD2L1	5.14	3.24e-38	0.000
5	ENSG00000140534	TICRR	5.14	2.64e-34	0.000
5	ENSG00000237649	KIFC1	5.02	3.82e-39	0.000
5	ENSG00000203760	CENPW	4.88	1.58e-30	0.000
5	ENSG00000105011	ASF1B	5.59	3.67e-29	0.001
5	ENSG00000035499	DEPDC1B	5	7.37e-29	0.001
5	ENSG00000158402	CDC25C	5.03	2.19e-28	0.004
5	ENSG00000135451	TROAP	5.81	1.05e-27	0.011
5	ENSG00000092853	CLSPN	4.79	1.87e-27	0.012
5	ENSG00000139734	DIAPH3	4.88	2.59e-27	0.014
5	ENSG00000164104	HMGB2	4.62	8.31e-27	0.019
5	ENSG00000138182	KIF20B	4.84	2.94e-25	0.046

Table A.12: Barrett's esophagus differentially expressed genes. The table displays the genes identified as significant (FDR <0.05) in differential expression analysis of the Barrett's single cell clusters. Genes are ordered by cluster and significance and the log2 fold-change in the cluster compared to the rest of the Barrett's biopsy single cell libraries is listed for each gene.

A.2.7 Differentially expressed genes in clusters from all QC-filtered cells

Table A.13 Differentially expressed genes in all QC-filtered cells.

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000132465	JCHAIN	7.33	2.68e-91	0.000
1	ENSG00000249119	MTND6P4	2.45	1.2e-117	0.000
1	ENSG00000232346	ENSG00000232346	2.42	7.88e-187	0.000
1	ENSG00000219928	ENSG00000219928	1.96	1.76e-151	0.000
1	ENSG00000181524	RPL24P4	1.96	3.8e-112	0.000
1	ENSG00000231767	ENSG00000231767	1.93	2.78e-151	0.000
1	ENSG00000218175	ENSG00000218175	1.86	4.56e-148	0.000
1	ENSG00000237991	RPL35P1	1.81	2.99e-105	0.000
1	ENSG00000180211	ENSG00000180211	1.8	4.92e-107	0.000
1	ENSG00000239559	RPL37P2	1.8	1.94e-91	0.000
1	ENSG00000224550	ENSG00000224550	1.77	4.82e-83	0.000
1	ENSG00000230897	RPS18P12	1.74	1.61e-95	0.000
1	ENSG00000225067	RPL23AP2	1.7	6.43e-102	0.000
1	ENSG00000226525	RPS7P10	1.67	1.04e-100	0.000
1	ENSG00000236801	RPL24P8	1.62	6.6e-99	0.000
1	ENSG00000146677	ENSG00000146677	1.61	1.47e-133	0.000
1	ENSG00000224631	ENSG00000224631	1.57	3.76e-128	0.000
1	ENSG00000178429	RPS3AP5	1.52	1.37e-107	0.000
1	ENSG00000228929	RPS13P2	1.44	2.64e-94	0.000
1	ENSG00000213598	ENSG00000213598	1.38	9.34e-123	0.000
1	ENSG00000230629	RPS23P8	1.33	2.44e-92	0.000
1	ENSG00000226084	ENSG00000226084	1.32	4.89e-82	0.000
1	ENSG00000214485	RPL7P1	1.31	6.05e-82	0.000
1	ENSG00000136810	TXN	1.28	1.13e-83	0.000
1	ENSG00000214535	RPS15AP1	1.77	9.76e-80	0.001
1	ENSG00000227615	ENSG00000227615	1.32	1.34e-80	0.001
1	ENSG00000235174	RPL39P3	1.33	8.7e-76	0.002
1	ENSG00000185641	ENSG00000185641	1.16	2.08e-75	0.002
1	ENSG00000219023	ENSG00000219023	1.71	8.78e-74	0.004
1	ENSG00000211899	IGHM	6.91	1.51e-72	0.005
1	ENSG00000211890	IGHA2	6.43	8.33e-72	0.006
1	ENSG00000234036	TXNP6	1.67	6.35e-70	0.009
1	ENSG00000223803	RPS20P14	1.68	1.21e-69	0.01
1	ENSG00000226221	RPL26P19	1.24	1.05e-69	0.01
1	ENSG00000075624	ACTB	-1.13	1.18e-88	0.000
1	ENSG00000167996	FTH1	-1.92	2.79e-115	0.000
1	ENSG00000167004	PDIA3	-2.06	1.2e-82	0.000
1	ENSG00000101439	CST3	-2.3	8.33e-90	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000263563	UBBP4	-2.41	2.57e-97	0.000
1	ENSG00000111057	KRT18	-3.18	1.25e-98	0.000
1	ENSG00000254285	KRT8P3	-3.54	3.89e-90	0.000
1	ENSG00000240668	KRT8P36	-3.83	2.59e-84	0.000
1	ENSG00000170421	KRT8	-4.22	2.8e-104	0.000
1	ENSG00000119888	EPCAM	-4.81	5.83e-98	0.000
1	ENSG00000175130	MARCKSL1	-5.33	7.94e-96	0.000
1	ENSG00000149260	CAPN5	-5.72	1.16e-85	0.000
1	ENSG00000135052	GOLM1	-5.83	7.13e-97	0.000
1	ENSG00000101076	HNF4A	-5.98	3.8e-111	0.000
1	ENSG00000112699	GMDS	-5.98	5.83e-123	0.000
1	ENSG00000145703	IQGAP2	-6.07	6.1e-107	0.000
1	ENSG00000162545	CAMK2N1	-6.09	2.15e-96	0.000
1	ENSG00000162896	PIGR	-6.15	2.37e-82	0.000
1	ENSG00000169562	GJB1	-6.26	8.19e-99	0.000
1	ENSG00000198643	FAM3D	-6.28	6.74e-86	0.000
1	ENSG00000267795	SMIM22	-6.56	2.31e-144	0.000
1	ENSG00000141448	GATA6	-6.66	2.81e-114	0.000
1	ENSG00000184012	TMPRSS2	-6.69	8.36e-93	0.000
1	ENSG00000137501	SYTL2	-6.78	1.7e-84	0.000
1	ENSG00000007062	PROM1	-6.79	1.42e-87	0.000
1	ENSG00000127324	TSPAN8	-6.83	1.27e-120	0.000
1	ENSG00000106541	AGR2	-6.86	2.1e-103	0.000
1	ENSG00000160862	AZGP1	-6.95	1.56e-84	0.000
1	ENSG00000173597	SULT1B1	-7.06	2.74e-87	0.000
1	ENSG00000173467	AGR3	-7.11	4.21e-107	0.000
1	ENSG00000103534	TMC5	-7.15	7.13e-117	0.000
1	ENSG00000204385	SLC44A4	-7.22	3.6e-102	0.000
1	ENSG00000095932	SMIM24	-7.27	2.23e-107	0.000
1	ENSG00000006611	USH1C	-7.27	1.37e-132	0.000
1	ENSG00000134398	ERN2	-7.33	5.33e-85	0.000
1	ENSG00000173702	MUC13	-7.33	7.84e-99	0.000
1	ENSG00000002726	AOC1	-7.36	1.75e-94	0.000
1	ENSG00000164266	SPINK1	-7.36	7.64e-94	0.000
1	ENSG00000115648	MLPH	-7.37	2.32e-84	0.000
1	ENSG00000131771	PPP1R1B	-7.5	7.26e-119	0.000
1	ENSG00000163993	S100P	-7.52	2.98e-82	0.000
1	ENSG00000116299	KIAA1324	-7.82	6.41e-124	0.000
1	ENSG00000171747	LGALS4	-7.97	8.42e-126	0.000
1	ENSG00000185499	MUC1	-7.97	6e-99	0.000
1	ENSG00000064787	BCAS1	-8.02	1.11e-111	0.000
1	ENSG00000117983	MUC5B	-8.04	6.62e-82	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000203697	CAPN8	-8.04	7.88e-91	0.000
1	ENSG00000160182	TFF1	-8.39	1.58e-82	0.000
1	ENSG00000066405	CLDN18	-8.43	3.77e-88	0.000
1	ENSG00000196188	CTSE	-8.64	2.03e-95	0.000
1	ENSG00000167658	EEF2	-1.65	5.57e-81	0.001
1	ENSG00000165092	ALDH1A1	-4.27	1.06e-76	0.001
1	ENSG00000120306	CYSTM1	-4.44	1.92e-80	0.001
1	ENSG00000162482	AKR7A3	-5.58	2.12e-79	0.001
1	ENSG00000166920	C15orf48	-6.14	2.01e-78	0.001
1	ENSG00000156966	B3GNT7	-6.8	1.63e-78	0.001
1	ENSG00000099812	MISP	-6.93	7.38e-80	0.001
1	ENSG00000019102	VSIG2	-7.39	3.21e-77	0.001
1	ENSG00000172757	CFL1	-1.16	5.39e-75	0.002
1	ENSG00000223361	FTH1P10	-2.47	3.28e-76	0.002
1	ENSG00000143153	ATP1B1	-2.7	4.32e-76	0.002
1	ENSG00000064651	SLC12A2	-3.59	1.2e-74	0.002
1	ENSG00000164749	HNF4G	-5.77	4.36e-75	0.002
1	ENSG00000134240	HMGCS2	-5.85	1.33e-74	0.002
1	ENSG00000180745	CLRN3	-5.89	6.54e-75	0.002
1	ENSG00000142484	TM4SF5	-6.5	1.41e-76	0.002
1	ENSG00000127831	VIL1	-6.52	1.31e-75	0.002
1	ENSG00000198203	SULT1C2	-7.59	2.99e-74	0.002
1	ENSG00000107159	CA9	-7.75	1.59e-75	0.002
1	ENSG00000109511	ANXA10	-7.86	1.84e-74	0.002
1	ENSG00000198758	EPS8L3	-6.47	3.56e-74	0.003
1	ENSG00000167183	PRR15L	-5.01	7.96e-74	0.004
1	ENSG00000144820	ADGRG7	-6.02	2.24e-73	0.004
1	ENSG00000012171	SEMA3B	-7	7.9e-74	0.004
1	ENSG00000064270	ATP2C2	-5.49	2.89e-72	0.005
1	ENSG00000139540	SLC39A5	-6.76	1.06e-72	0.005
1	ENSG00000185624	P4HB	-1.78	3.45e-71	0.006
1	ENSG00000219507	FTH1P8	-2.32	1.71e-70	0.006
1	ENSG00000081923	ATP8B1	-3.39	4.73e-72	0.006
1	ENSG00000103187	COTL1	-4.67	3.04e-71	0.006
1	ENSG00000080493	SLC4A4	-5.53	4.92e-72	0.006
1	ENSG00000101577	LPIN2	-5.75	1.44e-71	0.006
1	ENSG00000233041	PHGR1	-6.42	6.1e-72	0.006
1	ENSG00000197249	SERPINA1	-7.75	3.48e-71	0.006
1	ENSG00000160181	TFF2	-8.83	1.42e-70	0.006
1	ENSG00000214530	STARD10	-3.67	9.19e-70	0.009
1	ENSG00000113924	HGD	-5.75	1.4e-69	0.01
1	ENSG00000140391	TSPAN3	-2.72	2.22e-69	0.011

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
1	ENSG00000257743	MGAM2	-6.2	2.12e-69	0.011
1	ENSG00000088992	TESC	-6.7	1.91e-69	0.011
1	ENSG00000122176	FMOD	-6.94	1.83e-69	0.011
1	ENSG00000140416	TPM1	-3.46	6.7e-69	0.012
1	ENSG00000164237	CMBL	-4.83	2.84e-69	0.012
1	ENSG00000117525	F3	-6.19	5.75e-69	0.012
1	ENSG00000169994	MYO7B	-6.4	6.55e-69	0.012
1	ENSG00000169894	MUC3A	-6.78	2.14e-68	0.013
1	ENSG00000172270	BSG	-1.52	4.23e-67	0.024
1	ENSG00000204525	HLA-C	-1.7	8.39e-67	0.024
1	ENSG00000175538	KCNE3	-4.8	6.9e-67	0.024
1	ENSG00000163898	LIPH	-5.5	3.85e-67	0.024
1	ENSG00000166866	MYO1A	-6.8	2.97e-67	0.024
1	ENSG00000179913	B3GNT3	-6.38	9.2e-66	0.044
2	ENSG00000132911	NMUR2	5.75	1.28e-103	0.000
2	ENSG00000243709	LEFTY1	5.62	6.56e-110	0.000
2	ENSG00000105894	PTN	-7.28	3.76e-92	0.000
2	ENSG00000163331	DAPL1	-7.41	1.16e-93	0.000
2	ENSG00000171401	KRT13	-7.67	5.37e-91	0.000
2	ENSG00000230937	MIR205HG	-8.18	6.28e-98	0.000
2	ENSG00000178363	CALML3	-8.43	5.69e-95	0.000
2	ENSG00000186081	KRT5	-8.68	1.7e-95	0.000
2	ENSG00000163220	S100A9	-8.81	5.54e-98	0.000
2	ENSG00000167656	LY6D	-8.82	5.78e-100	0.000
2	ENSG00000196754	S100A2	-7.2	5.4e-85	0.001
2	ENSG00000137975	CLCA2	-8.07	4.18e-89	0.001
2	ENSG00000143546	S100A8	-8.08	7.29e-83	0.001
2	ENSG00000183347	GBP6	-8.45	2.83e-87	0.001
2	ENSG00000205420	KRT6A	-8.51	7.49e-87	0.001
2	ENSG00000140519	RHCG	-9.13	7.4e-87	0.001
2	ENSG00000134762	DSC3	-7.35	2.57e-80	0.003
2	ENSG00000124466	LYPD3	-8.07	1.97e-80	0.003
2	ENSG00000121552	CSTA	-6.84	9.66e-79	0.004
2	ENSG00000081277	PKP1	-7.37	1.13e-79	0.004
2	ENSG00000178934	LGALS7B	-6.53	1.46e-74	0.036
3	ENSG00000183607	GKN2	9.5	0.000	0.000
3	ENSG00000169605	GKN1	9.47	0.000	0.000
3	ENSG00000229859	PGA3	8.92	0.000	0.000
3	ENSG00000096088	PGC	8.75	3.38e-238	0.000
3	ENSG00000229183	PGA4	8.33	0.000	0.000
3	ENSG00000168631	DPCR1	7.66	1.04e-142	0.000
3	ENSG00000182333	LIPF	7.08	3.41e-122	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
3	ENSG00000125999	BPIFB1	7	1.81e-149	0.000
3	ENSG00000105641	SLC5A5	6.87	3.26e-115	0.000
3	ENSG00000163631	ALB	6.29	1.64e-191	0.000
3	ENSG00000167653	PSCA	6.18	1.31e-97	0.000
3	ENSG00000110195	FOLR1	6.16	7.02e-99	0.000
3	ENSG00000256713	PGA5	5.37	7.29e-182	0.000
3	ENSG00000159197	KCNE2	5.12	7.23e-88	0.000
3	ENSG00000186009	ATP4B	4.86	1.09e-100	0.000
3	ENSG00000197353	LYPD2	4.62	2.76e-95	0.000
3	ENSG00000247809	NR2F2-AS1	4.52	3.75e-90	0.000
3	ENSG00000174501	ANKRD36C	4.42	2.31e-89	0.000
3	ENSG00000205542	TMSB4X	1.77	2.66e-84	0.006
3	ENSG00000134812	GIF	4.97	4.06e-79	0.038
3	ENSG00000026508	CD44	-5.12	7.54e-79	0.041
4	ENSG00000229932	YWHAZP3	1.74	1.13e-80	0.005
4	ENSG00000173801	JUP	2.27	8.16e-79	0.01
4	ENSG00000175130	MARCKSL1	-8.16	1.32e-106	0.000
4	ENSG00000116299	KIAA1324	-8.35	1.14e-89	0.000
4	ENSG00000106541	AGR2	-8.84	2.77e-85	0.000
4	ENSG00000267795	SMIM22	-6.29	5.1e-82	0.002
4	ENSG00000171747	LGALS4	-8.79	4.99e-82	0.002
4	ENSG00000103534	TMC5	-7.46	2.68e-79	0.008
4	ENSG00000204385	SLC44A4	-7.81	3.16e-78	0.013
4	ENSG00000127324	TSPAN8	-7.67	4.7e-78	0.015
4	ENSG00000090382	LYZ	-9.36	1.37e-76	0.035
4	ENSG00000145703	IQGAP2	-6.7	2.08e-76	0.037
4	ENSG00000215182	MUC5AC	-11.5	5.18e-76	0.042
5	ENSG00000110244	APOA4	10.1	3.47e-268	0.000
5	ENSG00000114113	RBP2	9.41	0.000	0.000
5	ENSG00000084674	APOB	8.84	7.21e-209	0.000
5	ENSG00000160868	CYP3A4	8.8	1.04e-297	0.000
5	ENSG00000115850	LCT	8.66	2.69e-238	0.000
5	ENSG00000131142	CCL25	8.44	5.1e-240	0.000
5	ENSG00000091138	SLC26A3	7.95	6.66e-182	0.000
5	ENSG00000172689	MS4A10	7.72	4.51e-179	0.000
5	ENSG00000110245	APOC3	7.72	7.55e-131	0.000
5	ENSG00000138823	MTTP	7.68	3.48e-255	0.000
5	ENSG00000105398	SULT2A1	7.66	3.14e-303	0.000
5	ENSG00000163295	ALPI	7.65	3.03e-126	0.000
5	ENSG00000196549	MME	7.25	1.15e-187	0.000
5	ENSG00000163581	SLC2A2	7.08	3.27e-129	0.000
5	ENSG00000249948	GBA3	6.96	3.08e-176	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
5	ENSG00000060566	CREB3L3	6.76	2.9e-128	0.000
5	ENSG00000197888	UGT2B17	6.73	2.85e-165	0.000
5	ENSG00000136872	ALDOB	6.72	6.37e-111	0.000
5	ENSG00000007216	SLC13A2	6.7	2.12e-121	0.000
5	ENSG00000021488	SLC7A9	6.21	3.2e-122	0.000
5	ENSG00000080293	SCTR	6.11	7.78e-143	0.000
5	ENSG00000174885	NLRP6	6.08	6.92e-165	0.000
5	ENSG00000131482	G6PC	6.07	1.42e-140	0.000
5	ENSG00000156222	SLC28A1	6.04	7.27e-144	0.000
5	ENSG00000135220	UGT2A3	6.03	3.03e-111	0.000
5	ENSG00000244067	GSTA2	5.81	3.52e-137	0.000
5	ENSG00000145626	UGT3A1	5.42	1.96e-124	0.000
5	ENSG00000169906	S100G	5.35	1.31e-122	0.000
5	ENSG00000256162	SMLR1	5.29	5.89e-135	0.000
5	ENSG00000138308	PLA2G12B	5.26	4.76e-147	0.000
5	ENSG00000172023	REG1B	5.1	4.51e-170	0.000
5	ENSG00000182793	GSTA5	4.31	2.32e-129	0.000
5	ENSG00000088386	SLC15A1	6.56	6.5e-106	0.001
5	ENSG00000124253	PCK1	6.56	9.03e-107	0.001
5	ENSG00000122121	XPNPEP2	6.04	6.67e-94	0.001
5	ENSG00000135218	CD36	6.02	6.84e-105	0.001
5	ENSG00000100170	SLC5A1	5.91	3.87e-104	0.001
5	ENSG00000117834	SLC5A9	5.81	3.36e-92	0.001
5	ENSG00000138792	ENPEP	5.79	1.48e-104	0.001
5	ENSG00000175329	ISX	5.76	1.12e-95	0.001
5	ENSG00000143595	AQP10	5.64	1.84e-98	0.001
5	ENSG00000148942	SLC5A12	5.25	5.08e-92	0.001
5	ENSG00000165828	PRAP1	4.89	5.13e-93	0.001
5	ENSG00000214782	MS4A18	5.5	5.68e-91	0.002
5	ENSG00000188611	ASAH2	4.63	6.3e-91	0.002
5	ENSG00000197165	SULT1A2	4.37	2.62e-90	0.003
5	ENSG00000204978	ERICH4	5.19	2.45e-86	0.009
5	ENSG00000134612	FOLH1B	3.63	1.63e-85	0.012
5	ENSG00000186198	SLC51B	5.46	2.97e-84	0.023
5	ENSG00000162460	TMEM82	5.33	1.51e-83	0.03
5	ENSG00000146285	SCML4	5	2.31e-81	0.048
5	ENSG00000089356	FXYD3	-3.45	6.89e-103	0.001
5	ENSG00000163191	S100A11	-2.68	1.58e-83	0.03
6a	ENSG00000140287	HDC	9.5	5.42e-161	0.000
6a	ENSG00000118271	TTR	9.35	3.2e-229	0.000
6a	ENSG00000096395	MLN	8.67	2.59e-151	0.000
6a	ENSG00000109072	VTN	8.27	2.95e-137	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6a	ENSG00000162992	NEUROD1	8.26	2.43e-133	0.000
6a	ENSG00000157017	GHRL	8.24	1.03e-143	0.000
6a	ENSG00000079689	SCGN	7.98	2.7e-128	0.000
6a	ENSG00000100604	CHGA	7.93	6.79e-117	0.000
6a	ENSG00000163497	FEV	7.87	2.39e-127	0.000
6a	ENSG00000136698	CFC1	7.72	2.36e-175	0.000
6a	ENSG00000115263	GCG	7.58	1.97e-118	0.000
6a	ENSG00000177669	MBOAT4	7.53	9.12e-161	0.000
6a	ENSG00000163499	CRYBA2	7.24	1.35e-109	0.000
6a	ENSG00000139219	COL2A1	6.77	4.72e-110	0.000
6a	ENSG00000116774	OLFML3	7.66	4.13e-104	0.004
6a	ENSG00000147606	SLC26A7	6.15	1.63e-99	0.019
6a	ENSG00000070031	SCT	6.1	1.18e-98	0.022
6a	ENSG00000153002	CPB1	6.73	1.86e-92	0.045
6a	ENSG00000171951	SCG2	7.34	1.03e-89	0.046
6b	ENSG00000138755	CXCL9	8.08	9.46e-264	0.000
6b	ENSG00000173432	SAA1	7.48	7.13e-179	0.000
6b	ENSG00000131203	IDO1	7.04	2.48e-119	0.000
6b	ENSG00000206073	SERPINB4	6.85	1.5e-117	0.000
6b	ENSG00000143341	HMCN1	6.33	2.3e-122	0.000
6b	ENSG00000154451	GBP5	6.26	6.78e-115	0.000
6b	ENSG00000205076	LGALS7	6.15	0.000	0.000
6b	ENSG00000128422	KRT17	4.88	9.86e-131	0.000
6b	ENSG00000226025	ENSG00000226025	4.64	5.85e-90	0.039
6d	ENSG00000123560	PLP1	11.5	0.000	0.000
6d	ENSG00000104833	TUBB4A	11.4	0.000	0.000
6d	ENSG00000131095	GFAP	11.3	0.000	0.000
6d	ENSG00000105695	MAG	11	0.000	0.000
6d	ENSG00000107317	PTGDS	10.9	0.000	0.000
6d	ENSG00000189058	APOD	10.7	0.000	0.000
6d	ENSG00000168309	FAM107A	10.6	0.000	0.000
6d	ENSG00000160307	S100B	10.2	0.000	0.000
6d	ENSG00000115468	EFHD1	10	0.000	0.000
6d	ENSG00000118785	SPP1	9.92	0.000	0.000
6d	ENSG00000109846	CRYAB	9.78	0.000	0.000
6d	ENSG00000108387	SEPT4	9.7	0.000	0.000
6d	ENSG00000018625	ATP1A2	9.69	0.000	0.000
6d	ENSG00000166963	MAP1A	9.52	1.38e-302	0.000
6d	ENSG00000070808	CAMK2A	9.48	0.000	0.000
6d	ENSG00000184144	CNTN2	9.47	0.000	0.000
6d	ENSG00000087250	MT3	9.35	0.000	0.000
6d	ENSG00000204655	MOG	9.33	0.000	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000160781	PAQR6	9.29	8.75e-240	0.000
6d	ENSG00000172020	GAP43	9.28	0.000	0.000
6d	ENSG00000150625	GPM6A	9.27	0.000	0.000
6d	ENSG00000168314	MOBP	9.26	0.000	0.000
6d	ENSG00000013297	CLDN11	9.21	0.000	0.000
6d	ENSG00000152377	SPOCK1	9.2	0.000	0.000
6d	ENSG00000129244	ATP1B2	9.2	0.000	0.000
6d	ENSG00000010282	HHATL	9.18	0.000	0.000
6d	ENSG00000149557	FEZ1	9.15	5.99e-283	0.000
6d	ENSG00000015592	STMN4	9.15	0.000	0.000
6d	ENSG00000110436	SLC1A2	9.11	0.000	0.000
6d	ENSG00000154096	THY1	9.05	0.000	0.000
6d	ENSG00000105696	TMEM59L	8.97	0.000	0.000
6d	ENSG00000074317	SNCB	8.92	0.000	0.000
6d	ENSG00000130203	APOE	8.87	7.11e-249	0.000
6d	ENSG00000155980	KIF5A	8.81	1.01e-258	0.000
6d	ENSG00000244734	HBB	8.77	0.000	0.000
6d	ENSG00000166448	TMEM130	8.76	2.2e-274	0.000
6d	ENSG00000171885	AQP4	8.66	0.000	0.000
6d	ENSG00000132692	BCAN	8.64	0.000	0.000
6d	ENSG00000135472	FAIM2	8.63	7.1e-219	0.000
6d	ENSG00000124126	PREX1	8.59	1.72e-261	0.000
6d	ENSG00000167641	PPP1R14A	8.56	0.000	0.000
6d	ENSG00000146122	DAAM2	8.5	0.000	0.000
6d	ENSG00000187957	DNER	8.5	0.000	0.000
6d	ENSG00000100427	MLC1	8.49	1.22e-274	0.000
6d	ENSG00000184258	CDR1	8.46	2.39e-268	0.000
6d	ENSG00000184221	OLIG1	8.44	0.000	0.000
6d	ENSG00000177807	KCNJ10	8.41	0.000	0.000
6d	ENSG00000197430	OPALIN	8.4	0.000	0.000
6d	ENSG00000141540	TTYH2	8.36	3e-210	0.000
6d	ENSG00000104888	SLC17A7	8.36	0.000	0.000
6d	ENSG00000135744	AGT	8.36	6.7e-156	0.000
6d	ENSG00000105270	CLIP3	8.35	2.92e-215	0.000
6d	ENSG00000136541	ERMN	8.31	9.61e-275	0.000
6d	ENSG00000105290	APLP1	8.3	8.61e-148	0.000
6d	ENSG00000162188	GNG3	8.3	0.000	0.000
6d	ENSG00000125462	C1orf61	8.28	7.39e-301	0.000
6d	ENSG00000154277	UCHL1	8.25	2.41e-193	0.000
6d	ENSG00000132718	SYT11	8.24	2.42e-182	0.000
6d	ENSG00000046653	GPM6B	8.24	6.5e-226	0.000
6d	ENSG00000149294	NCAM1	8.23	2.03e-183	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000170075	GPR37L1	8.23	8.54e-227	0.000
6d	ENSG00000169862	CTNND2	8.22	0.000	0.000
6d	ENSG00000167654	ATCAY	8.21	0.000	0.000
6d	ENSG00000184524	CEND1	8.17	0.000	0.000
6d	ENSG00000167755	KLK6	8.11	6.56e-182	0.000
6d	ENSG00000067715	SYT1	8.09	6.21e-145	0.000
6d	ENSG00000144834	TAGLN3	8.08	4.04e-238	0.000
6d	ENSG00000167614	TTYH1	8.08	1.15e-172	0.000
6d	ENSG00000197971	MBP	8.08	0.000	0.000
6d	ENSG00000124507	PACSIN1	8.04	1.61e-247	0.000
6d	ENSG00000137726	FXYD6	8.01	2.66e-235	0.000
6d	ENSG00000166091	CMTM5	8.01	0.000	0.000
6d	ENSG00000165478	HEPACAM	8	0.000	0.000
6d	ENSG00000157851	DPYSL5	7.98	2.99e-256	0.000
6d	ENSG00000008056	SYN1	7.95	0.000	0.000
6d	ENSG00000101265	RASSF2	7.95	5.08e-304	0.000
6d	ENSG00000170775	GPR37	7.94	3.21e-180	0.000
6d	ENSG00000155886	SLC24A2	7.94	0.000	0.000
6d	ENSG00000134769	DTNA	7.88	0.000	0.000
6d	ENSG00000128656	CHN1	7.86	9.79e-202	0.000
6d	ENSG00000159164	SV2A	7.84	0.000	0.000
6d	ENSG00000242808	SOX2-OT	7.82	1.13e-201	0.000
6d	ENSG00000196104	SPOCK3	7.79	0.000	0.000
6d	ENSG00000164929	BAALC	7.79	5.45e-215	0.000
6d	ENSG00000089169	RPH3A	7.72	0.000	0.000
6d	ENSG00000132702	HAPLN2	7.69	0.000	0.000
6d	ENSG00000100167	SEPT3	7.69	6.52e-184	0.000
6d	ENSG00000155849	ELMO1	7.66	3.45e-163	0.000
6d	ENSG00000213760	ATP6V1G2	7.65	1.17e-268	0.000
6d	ENSG00000196338	NLGN3	7.61	2.4e-189	0.000
6d	ENSG00000130294	KIF1A	7.61	4.01e-130	0.000
6d	ENSG00000007237	GAS7	7.6	1.72e-202	0.000
6d	ENSG00000164076	CAMKV	7.57	1.08e-281	0.000
6d	ENSG00000188536	HBA2	7.56	0.000	0.000
6d	ENSG00000078053	AMPH	7.55	0.000	0.000
6d	ENSG00000113578	FGF1	7.52	1.39e-175	0.000
6d	ENSG00000185046	ANKS1B	7.47	1.12e-209	0.000
6d	ENSG00000198822	GRM3	7.47	2.46e-194	0.000
6d	ENSG00000147588	PMP2	7.46	0.000	0.000
6d	ENSG00000105409	ATP1A3	7.46	0.000	0.000
6d	ENSG00000165566	AMER2	7.46	1.13e-308	0.000
6d	ENSG00000021300	PLEKHB1	7.46	5.79e-129	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000145284	SCD5	7.46	5.43e-154	0.000
6d	ENSG00000034053	APBA2	7.45	1.33e-243	0.000
6d	ENSG00000075043	KCNQ2	7.44	0.000	0.000
6d	ENSG00000182902	SLC25A18	7.43	1.06e-215	0.000
6d	ENSG00000157782	CABP1	7.41	0.000	0.000
6d	ENSG00000169006	NTSR2	7.41	1.33e-197	0.000
6d	ENSG00000100146	SOX10	7.38	1.57e-214	0.000
6d	ENSG00000152092	ASTN1	7.38	0.000	0.000
6d	ENSG00000164089	ETNPPL	7.37	0.000	0.000
6d	ENSG00000131711	MAP1B	7.36	8.75e-123	0.000
6d	ENSG00000078328	RBFOX1	7.36	0.000	0.000
6d	ENSG00000130287	NCAN	7.35	0.000	0.000
6d	ENSG00000066032	CTNNA2	7.34	0.000	0.000
6d	ENSG00000104722	NEFM	7.33	4.4e-142	0.000
6d	ENSG00000169554	ZEB2	7.31	3.55e-184	0.000
6d	ENSG00000135636	DYSF	7.3	2.48e-187	0.000
6d	ENSG00000198597	ZNF536	7.29	1.09e-212	0.000
6d	ENSG00000116675	DNAJC6	7.26	2.87e-155	0.000
6d	ENSG00000101445	PPP1R16B	7.25	2.92e-139	0.000
6d	ENSG00000158859	ADAMTS4	7.24	7.42e-186	0.000
6d	ENSG00000136274	NACAD	7.22	4.36e-135	0.000
6d	ENSG00000087495	PHACTR3	7.2	1.99e-182	0.000
6d	ENSG00000168481	LGI3	7.18	6.42e-157	0.000
6d	ENSG00000154146	NRGN	7.16	0.000	0.000
6d	ENSG00000104435	STMN2	7.16	2.63e-155	0.000
6d	ENSG00000168913	ENHO	7.15	0.000	0.000
6d	ENSG00000214548	MEG3	7.14	7.16e-168	0.000
6d	ENSG00000131386	GALNT15	7.14	4.3e-293	0.000
6d	ENSG00000105649	RAB3A	7.12	7.43e-169	0.000
6d	ENSG00000101463	SYNDIG1	7.1	7.01e-152	0.000
6d	ENSG00000103647	CORO2B	7.09	2.18e-249	0.000
6d	ENSG00000139970	RTN1	7.06	2.63e-110	0.000
6d	ENSG00000022355	GABRA1	7.05	0.000	0.000
6d	ENSG00000069424	KCNAB2	7.05	9.98e-152	0.000
6d	ENSG00000109107	ALDOC	7.04	2.68e-128	0.000
6d	ENSG00000173267	SNCG	7.03	6.22e-145	0.000
6d	ENSG00000116983	HPCAL4	7.03	0.000	0.000
6d	ENSG00000092051	JPH4	7.01	3.38e-245	0.000
6d	ENSG00000171246	NPTX1	7.01	5.44e-145	0.000
6d	ENSG00000065609	SNAP91	7	0.000	0.000
6d	ENSG00000157087	ATP2B2	6.99	2.67e-210	0.000
6d	ENSG00000092096	SLC22A17	6.99	5.18e-108	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000134817	APLNR	6.98	2.3e-104	0.000
6d	ENSG00000185565	LSAMP	6.97	1.11e-196	0.000
6d	ENSG00000180929	GPR62	6.97	4.85e-128	0.000
6d	ENSG00000127561	SYNGR3	6.95	7.85e-141	0.000
6d	ENSG00000006116	CACNG3	6.93	2.93e-273	0.000
6d	ENSG00000130558	OLFM1	6.93	7.33e-116	0.000
6d	ENSG00000182667	NTM	6.92	0.000	0.000
6d	ENSG00000186310	NAP1L3	6.91	5.01e-312	0.000
6d	ENSG00000179915	NRXN1	6.9	1.3e-238	0.000
6d	ENSG00000153902	LGI4	6.89	1.74e-175	0.000
6d	ENSG00000078549	ADCYAP1R1	6.89	4.63e-131	0.000
6d	ENSG00000154917	RAB6B	6.89	5.15e-133	0.000
6d	ENSG00000081189	MEF2C	6.87	1.84e-176	0.000
6d	ENSG00000159189	C1QC	6.86	1.42e-272	0.000
6d	ENSG00000106976	DNM1	6.85	6.52e-169	0.000
6d	ENSG00000169085	C8orf46	6.85	1.99e-132	0.000
6d	ENSG00000084453	SLCO1A2	6.84	4.28e-295	0.000
6d	ENSG00000180834	MAP6D1	6.84	2.3e-116	0.000
6d	ENSG00000107331	ABCA2	6.83	8.62e-126	0.000
6d	ENSG00000134594	RAB33A	6.82	1.01e-248	0.000
6d	ENSG00000109906	ZBTB16	6.82	7.1e-135	0.000
6d	ENSG00000167371	PRRT2	6.8	3.25e-174	0.000
6d	ENSG00000075340	ADD2	6.8	1.39e-213	0.000
6d	ENSG00000182013	PNMAL1	6.8	3.38e-188	0.000
6d	ENSG00000165424	ZCCHC24	6.79	5.52e-123	0.000
6d	ENSG00000145920	CPLX2	6.77	5.28e-101	0.000
6d	ENSG00000100285	NEFH	6.77	8.69e-130	0.000
6d	ENSG00000206172	HBA1	6.76	0.000	0.000
6d	ENSG00000197106	SLC6A17	6.76	7.54e-173	0.000
6d	ENSG00000141469	SLC14A1	6.75	1.31e-149	0.000
6d	ENSG00000038427	VCAN	6.74	5.17e-142	0.000
6d	ENSG00000058335	RASGRF1	6.72	4.85e-201	0.000
6d	ENSG00000184905	TCEAL2	6.72	2.23e-144	0.000
6d	ENSG00000172348	RCAN2	6.72	1.5e-128	0.000
6d	ENSG00000181418	DDN	6.71	1.17e-214	0.000
6d	ENSG00000130643	CALY	6.71	2.98e-126	0.000
6d	ENSG00000138311	ZNF365	6.71	4.49e-263	0.000
6d	ENSG00000173372	C1QA	6.69	1.57e-274	0.000
6d	ENSG00000133069	TMCC2	6.69	9.03e-121	0.000
6d	ENSG00000177108	ZDHHC22	6.68	4.45e-155	0.000
6d	ENSG00000198892	SHISA4	6.68	1.41e-140	0.000
6d	ENSG00000122733	PHF24	6.68	5.06e-146	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000132563	REEP2	6.68	4.64e-129	0.000
6d	ENSG00000134824	FADS2	6.67	8.96e-129	0.000
6d	ENSG00000237949	LINC00844	6.66	0.000	0.000
6d	ENSG00000171368	TPPP	6.65	1.8e-119	0.000
6d	ENSG00000187094	CCK	6.64	4.49e-102	0.000
6d	ENSG00000168490	PHYHIP	6.63	6.08e-103	0.000
6d	ENSG00000162706	CADM3	6.63	0.000	0.000
6d	ENSG00000081479	LRP2	6.62	1.07e-151	0.000
6d	ENSG00000117016	RIMS3	6.61	2.69e-117	0.000
6d	ENSG00000175874	CREG2	6.61	1.87e-170	0.000
6d	ENSG00000182916	TCEAL7	6.61	0.000	0.000
6d	ENSG00000173369	C1QB	6.6	6.65e-213	0.000
6d	ENSG00000185736	ADARB2	6.57	1.4e-168	0.000
6d	ENSG00000166501	PRKCB	6.56	8.85e-128	0.000
6d	ENSG00000182771	GRID1	6.55	5.71e-175	0.000
6d	ENSG00000165868	HSPA12A	6.55	1.35e-114	0.000
6d	ENSG00000144230	GPR17	6.53	9.76e-154	0.000
6d	ENSG00000188783	PRELP	6.52	1.52e-117	0.000
6d	ENSG00000103740	ACSBG1	6.52	4.23e-180	0.000
6d	ENSG00000121653	MAPK8IP1	6.52	1.73e-130	0.000
6d	ENSG00000005513	SOX8	6.51	2.12e-105	0.000
6d	ENSG00000205927	OLIG2	6.5	8.48e-263	0.000
6d	ENSG00000164176	EDIL3	6.5	2.42e-119	0.000
6d	ENSG00000167281	RBFOX3	6.46	9.9e-171	0.000
6d	ENSG00000137285	TUBB2B	6.46	3.24e-264	0.000
6d	ENSG00000152969	JAKMIP1	6.46	8.91e-168	0.000
6d	ENSG00000166257	SCN3B	6.42	1.34e-106	0.000
6d	ENSG00000172247	C1QTNF4	6.42	1.81e-173	0.000
6d	ENSG00000196361	ELAVL3	6.39	0.000	0.000
6d	ENSG00000077080	ACTL6B	6.39	8.16e-239	0.000
6d	ENSG00000183837	PNMA3	6.39	2.39e-120	0.000
6d	ENSG00000185818	NAT8L	6.39	3.52e-101	0.000
6d	ENSG00000149575	SCN2B	6.38	4.49e-122	0.000
6d	ENSG00000158560	DYNC1I1	6.38	7.15e-110	0.000
6d	ENSG00000100095	SEZ6L	6.38	1.03e-233	0.000
6d	ENSG00000126861	OMG	6.37	4.37e-107	0.000
6d	ENSG00000087258	GNAO1	6.37	1.64e-102	0.000
6d	ENSG00000068976	PYGM	6.36	1.32e-152	0.000
6d	ENSG00000139200	PIANP	6.36	1.26e-237	0.000
6d	ENSG00000204071	TCEAL6	6.36	0.000	0.000
6d	ENSG00000136928	GABBR2	6.36	5.13e-116	0.000
6d	ENSG00000091513	TF	6.35	5.42e-131	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000164303	ENPP6	6.32	2.2e-155	0.000
6d	ENSG00000061918	GUCY1B3	6.31	1.91e-189	0.000
6d	ENSG00000036530	CYP46A1	6.3	9.14e-141	0.000
6d	ENSG00000157152	SYN2	6.3	3.45e-102	0.000
6d	ENSG00000172995	ARPP21	6.3	1.34e-226	0.000
6d	ENSG00000106624	AEBP1	6.28	1.85e-119	0.000
6d	ENSG00000092964	DPYSL2	6.28	5.08e-117	0.000
6d	ENSG00000101489	CELF4	6.28	1.19e-117	0.000
6d	ENSG00000184156	KCNQ3	6.26	2.92e-127	0.000
6d	ENSG00000221946	FXYD7	6.25	0.000	0.000
6d	ENSG00000103449	SALL1	6.25	4.15e-146	0.000
6d	ENSG00000185274	WBSCR17	6.24	3.79e-131	0.000
6d	ENSG00000204278	TMEM235	6.24	2.13e-151	0.000
6d	ENSG00000007264	MATK	6.24	2.65e-115	0.000
6d	ENSG00000164106	SCRG1	6.24	1.84e-144	0.000
6d	ENSG00000122756	CNTFR	6.23	9.38e-117	0.000
6d	ENSG00000174521	TTC9B	6.22	7.22e-220	0.000
6d	ENSG00000105255	FSD1	6.22	2.98e-124	0.000
6d	ENSG00000186417	GLDN	6.21	2.91e-117	0.000
6d	ENSG00000123360	PDE1B	6.2	2.73e-125	0.000
6d	ENSG00000135643	KCNMB4	6.19	1.63e-106	0.000
6d	ENSG00000187730	GABRD	6.17	4.3e-108	0.000
6d	ENSG00000166165	CKB	6.17	3.45e-127	0.000
6d	ENSG00000152954	NRSN1	6.16	2.03e-154	0.000
6d	ENSG00000223802	CERS1	6.15	2.89e-149	0.000
6d	ENSG00000110786	PTPN5	6.14	4.48e-109	0.000
6d	ENSG00000156475	PPP2R2B	6.14	4.64e-110	0.000
6d	ENSG00000166925	TSC22D4	6.14	1.04e-168	0.000
6d	ENSG00000109956	B3GAT1	6.13	1.09e-128	0.000
6d	ENSG00000154654	NCAM2	6.13	1.61e-163	0.000
6d	ENSG00000171798	KNDC1	6.12	1.42e-133	0.000
6d	ENSG00000161896	IP6K3	6.09	5.16e-132	0.000
6d	ENSG00000077522	ACTN2	6.09	2.37e-189	0.000
6d	ENSG00000185985	SLITRK2	6.09	1.98e-164	0.000
6d	ENSG00000008118	CAMK1G	6.07	1.62e-118	0.000
6d	ENSG00000131409	LRRC4B	6.07	1.97e-129	0.000
6d	ENSG00000120251	GRIA2	6.07	1.63e-114	0.000
6d	ENSG00000197177	ADGRA1	6.07	1.35e-109	0.000
6d	ENSG00000224924	LINC00320	6.06	1.79e-199	0.000
6d	ENSG00000114279	FGF12	6.06	2.8e-175	0.000
6d	ENSG00000138650	PCDH10	6.05	4.91e-139	0.000
6d	ENSG00000163288	GABRB1	6.04	1.35e-152	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000152910	CNTNAP4	6.04	6.83e-214	0.000
6d	ENSG00000186487	MYT1L	6.02	1.08e-141	0.000
6d	ENSG00000186732	MPPED1	6.02	1.24e-109	0.000
6d	ENSG00000197380	DACT3	5.99	3.39e-212	0.000
6d	ENSG00000167971	CASKIN1	5.96	7.81e-148	0.000
6d	ENSG00000149927	DOC2A	5.96	2.67e-121	0.000
6d	ENSG00000113327	GABRG2	5.95	2.29e-233	0.000
6d	ENSG00000227544	ENSG00000227544	5.95	2.8e-172	0.000
6d	ENSG00000149970	CNKSR2	5.95	3.3e-108	0.000
6d	ENSG00000182870	GALNT9	5.95	4.74e-117	0.000
6d	ENSG00000173786	CNP	5.95	1.11e-237	0.000
6d	ENSG00000176381	PRR18	5.95	2.72e-176	0.000
6d	ENSG00000146592	CREB5	5.94	1.59e-131	0.000
6d	ENSG00000067842	ATP2B3	5.94	5.01e-199	0.000
6d	ENSG00000101198	NKAIN4	5.93	1.99e-119	0.000
6d	ENSG00000167552	TUBA1A	5.92	9.54e-251	0.000
6d	ENSG00000053438	NNAT	5.92	1.18e-106	0.000
6d	ENSG00000198910	L1CAM	5.89	5.06e-235	0.000
6d	ENSG00000136535	TBR1	5.87	1.59e-132	0.000
6d	ENSG00000160539	PLPP7	5.85	4.27e-108	0.000
6d	ENSG00000122585	NPY	5.85	6.53e-285	0.000
6d	ENSG00000020129	NCDN	5.84	1.42e-112	0.000
6d	ENSG00000223609	HBD	5.84	5.33e-311	0.000
6d	ENSG00000250510	GPR162	5.84	4.24e-172	0.000
6d	ENSG00000072952	MRVI1	5.83	1.49e-117	0.000
6d	ENSG00000145423	SFRP2	5.83	3.33e-118	0.000
6d	ENSG00000117154	IGSF21	5.83	2.19e-133	0.000
6d	ENSG00000170558	CDH2	5.81	4.91e-180	0.000
6d	ENSG00000121769	FABP3	5.81	1.38e-204	0.000
6d	ENSG00000167889	MGAT5B	5.8	1.96e-110	0.000
6d	ENSG00000106689	LHX2	5.79	1.21e-164	0.000
6d	ENSG00000244274	DBNDD2	5.78	9.30e-217	0.000
6d	ENSG00000139220	PPFIA2	5.77	9.18e-129	0.000
6d	ENSG00000159784	FAM131B	5.76	4.09e-114	0.000
6d	ENSG00000116254	CHD5	5.76	1.89e-133	0.000
6d	ENSG00000168993	CPLX1	5.76	2.9e-103	0.000
6d	ENSG00000204065	TCEAL5	5.75	3.68e-301	0.000
6d	ENSG00000225465	RFPL1S	5.75	1.74e-146	0.000
6d	ENSG00000214575	CPEB1	5.75	1.06e-143	0.000
6d	ENSG00000170419	VSTM2A	5.7	4.84e-203	0.000
6d	ENSG00000145794	MEGF10	5.7	1.84e-129	0.000
6d	ENSG00000183715	OPCML	5.69	3.28e-166	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000207955	MIR219A2	5.68	3.22e-228	0.000
6d	ENSG00000165795	NDRG2	5.64	1.46e-170	0.000
6d	ENSG00000185742	C11orf87	5.64	1.38e-144	0.000
6d	ENSG00000136160	EDNRB	5.63	2.56e-143	0.000
6d	ENSG00000176884	GRIN1	5.61	9.63e-152	0.000
6d	ENSG00000122254	HS3ST2	5.59	1.05e-110	0.000
6d	ENSG00000101134	DOK5	5.59	6.57e-106	0.000
6d	ENSG00000135298	ADGRB3	5.57	1.27e-159	0.000
6d	ENSG00000183888	C1orf64	5.56	7.66e-115	0.000
6d	ENSG00000135439	AGAP2	5.55	3.54e-122	0.000
6d	ENSG00000103710	RASL12	5.55	2.86e-106	0.000
6d	ENSG00000170579	DLGAP1	5.54	3.23e-105	0.000
6d	ENSG00000166482	MFAP4	5.53	6.07e-103	0.000
6d	ENSG00000011600	TYROBP	5.51	2.49e-114	0.000
6d	ENSG00000145721	LIX1	5.51	1.45e-138	0.000
6d	ENSG00000011677	GABRA3	5.46	1.81e-104	0.000
6d	ENSG00000162734	PEA15	5.45	2.71e-112	0.000
6d	ENSG00000148482	SLC39A12	5.45	2.49e-107	0.000
6d	ENSG00000142408	CACNG8	5.44	2.27e-101	0.000
6d	ENSG00000110076	NRXN2	5.42	4.83e-135	0.000
6d	ENSG00000170390	DCLK2	5.41	5.2e-111	0.000
6d	ENSG00000166428	PLD4	5.4	1.02e-117	0.000
6d	ENSG00000203805	PLPP4	5.35	1.2e-110	0.000
6d	ENSG00000166006	KCNC2	5.32	1.63e-116	0.000
6d	ENSG00000171724	VAT1L	5.26	2.06e-103	0.000
6d	ENSG00000156076	WIF1	5.22	3.19e-106	0.000
6d	ENSG00000102468	HTR2A	5.21	2.07e-114	0.000
6d	ENSG00000005249	PRKAR2B	5.17	2.15e-115	0.000
6d	ENSG00000198914	POU3F3	5.14	2.08e-163	0.000
6d	ENSG00000141668	CBLN2	5.12	3.53e-110	0.000
6d	ENSG00000110675	ELMOD1	5.08	9.29e-148	0.000
6d	ENSG00000109738	GLRB	5.05	4.92e-104	0.000
6d	ENSG00000141750	STAC2	5.04	4.74e-125	0.000
6d	ENSG00000176165	FOXG1	5	2.6e-158	0.000
6d	ENSG00000139155	SLCO1C1	4.92	1.17e-105	0.000
6d	ENSG00000224795	NTM-AS1	4.88	2.11e-122	0.000
6d	ENSG00000163285	GABRG1	4.88	5.3e-110	0.000
6d	ENSG00000164434	FABP7	4.77	5.37e-139	0.000
6d	ENSG00000266964	FXYD1	4.74	2.93e-198	0.000
6d	ENSG00000263146	ENSG00000263146	4.69	1.05e-101	0.000
6d	ENSG00000233639	LINC01158	4.68	1.08e-122	0.000
6d	ENSG00000170370	EMX2	4.47	2.54e-125	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000124251	TP53TG5	4.39	2.46e-123	0.000
6d	ENSG00000205696	ADARB2-AS1	4.35	2.72e-116	0.000
6d	ENSG00000159176	CSRP1	4.11	3.25e-103	0.000
6d	ENSG00000121905	HPCA	4.09	3.87e-130	0.000
6d	ENSG00000100362	PVALB	4.08	3.41e-102	0.000
6d	ENSG00000260001	TGFBR3L	4.02	3.67e-104	0.000
6d	ENSG00000124785	NRN1	6.12	3.67e-100	0.001
6d	ENSG00000114757	PEX5L	5.92	1.73e-100	0.001
6d	ENSG00000178235	SLITRK1	5.56	1.72e-100	0.001
6d	ENSG00000004776	HSPB6	5	1.79e-100	0.001
6d	ENSG00000126878	AIF1L	6.66	9.99e-99	0.003
6d	ENSG00000102760	RGCC	6.59	1.17e-99	0.003
6d	ENSG00000108309	RUNDC3A	6.5	7.38e-100	0.003
6d	ENSG00000119865	CNRIP1	6.28	1.05e-97	0.003
6d	ENSG00000150510	FAM124A	6.18	7.98e-100	0.003
6d	ENSG00000135519	KCNH3	6.14	8.25e-99	0.003
6d	ENSG00000186868	MAPT	6.08	1.08e-97	0.003
6d	ENSG00000099365	STX1B	6	2.06e-98	0.003
6d	ENSG00000133083	DCLK1	5.78	2.38e-98	0.003
6d	ENSG00000108231	LGI1	5.43	1.84e-99	0.003
6d	ENSG00000162426	SLC45A1	5.42	1.66e-99	0.003
6d	ENSG00000196376	SLC35F1	5.01	2.96e-98	0.003
6d	ENSG00000152578	GRIA4	4.93	7.47e-97	0.005
6d	ENSG00000149090	PAMR1	5.66	2.12e-96	0.006
6d	ENSG00000144847	IGSF11	5.58	2.64e-96	0.006
6d	ENSG00000166033	HTRA1	6.67	9.61e-96	0.007
6d	ENSG00000102230	PCYT1B	5.74	3.29e-96	0.007
6d	ENSG00000154118	JPH3	6.28	2.85e-95	0.008
6d	ENSG00000150867	PIP4K2A	5.83	1.52e-95	0.008
6d	ENSG00000172137	CALB2	4.69	8.09e-95	0.008
6d	ENSG00000167191	GPRC5B	6.42	2.17e-94	0.01
6d	ENSG00000099194	SCD	5.17	2.52e-94	0.011
6d	ENSG00000186479	RGS7BP	4.94	5.71e-94	0.013
6d	ENSG00000129009	ISLR	5.86	8.36e-94	0.014
6d	ENSG00000146005	PSD2	6.06	2.77e-93	0.016
6d	ENSG00000101098	RIMS4	5.57	2.28e-93	0.016
6d	ENSG00000197457	STMN3	6.08	7.31e-93	0.021
6d	ENSG00000154493	C10orf90	5.5	1.73e-92	0.022
6d	ENSG00000155659	VSIG4	5.42	2.5e-92	0.023
6d	ENSG00000129596	CDO1	5.2	3.52e-92	0.023
6d	ENSG00000152137	HSPB8	6.52	4.19e-92	0.024
6d	ENSG00000117594	HSD11B1	4.79	2.16e-91	0.027

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000145832	SLC25A48	5.94	3.32e-91	0.028
6d	ENSG00000125869	LAMP5	5.31	2.88e-91	0.028
6d	ENSG00000164530	PI16	5.66	4.45e-91	0.029
6d	ENSG00000187135	VSTM2B	4.9	5.28e-91	0.029
6d	ENSG00000176533	GNG7	6.08	8.07e-91	0.032
6d	ENSG00000058404	CAMK2B	6.52	5.27e-90	0.033
6d	ENSG00000091129	NRCAM	6.14	3.53e-90	0.033
6d	ENSG00000124140	SLC12A5	6.08	6.68e-90	0.034
6d	ENSG00000162374	ELAVL4	5.7	6.69e-90	0.034
6d	ENSG00000145864	GABRB2	5.34	7.69e-90	0.034
6d	ENSG00000221890	NPTXR	5.89	5.82e-89	0.036
6d	ENSG00000141639	MAPK4	5.84	4.76e-89	0.036
6d	ENSG00000073670	ADAM11	5.49	5.68e-89	0.036
6d	ENSG00000012124	CD22	5.17	4.28e-89	0.036
6d	ENSG00000120645	IQSEC3	5.64	1.06e-88	0.037
6d	ENSG00000129990	SYT5	5.43	1.49e-88	0.037
6d	ENSG00000151025	GPR158	5.04	1.4e-88	0.037
6d	ENSG00000063180	CA11	5.5	3.63e-88	0.038
6d	ENSG00000166923	GREM1	5.26	2.25e-88	0.038
6d	ENSG00000140538	NTRK3	6.06	2.27e-87	0.039
6d	ENSG00000107130	NCS1	5.76	9.49e-88	0.039
6d	ENSG00000179242	CDH4	5.23	2.37e-87	0.039
6d	ENSG00000182132	KCNIP1	5.35	1.05e-86	0.041
6d	ENSG00000131097	HIGD1B	4.78	4.41e-87	0.041
6d	ENSG00000161082	CELF5	4.77	7.58e-87	0.041
6d	ENSG00000107105	ELAVL2	4.62	1.4e-86	0.042
6d	ENSG00000149633	KIAA1755	6.37	3.71e-86	0.043
6d	ENSG00000118946	PCDH17	5.69	3.91e-86	0.043
6d	ENSG00000104967	NOVA2	5.15	2.02e-86	0.043
6d	ENSG00000224189	HAGLR	5.24	4.38e-86	0.044
6d	ENSG00000117115	PADI2	6.19	4.93e-86	0.045
6d	ENSG00000126803	HSPA2	5.46	4.78e-86	0.045
6d	ENSG00000155511	GRIA1	5.1	6.38e-86	0.045
6d	ENSG00000059915	PSD	4.98	8.02e-86	0.045
6d	ENSG00000107518	ATRNL1	4.79	1.27e-85	0.045
6d	ENSG00000120885	CLU	4.78	1.12e-85	0.045
6d	ENSG00000197746	PSAP	2.8	9.36e-86	0.045
6d	ENSG00000198668	CALM1	2.22	7.27e-86	0.045
6d	ENSG00000109099	PMP22	6.59	1.57e-84	0.046
6d	ENSG00000111674	ENO2	6.47	1.15e-83	0.046
6d	ENSG00000179292	TMEM151A	6.07	4.5e-84	0.046
6d	ENSG00000166780	C16orf45	5.98	9.79e-83	0.046

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
6d	ENSG00000111783	RFX4	5.97	3.21e-83	0.046
6d	ENSG00000130226	DPP6	5.91	9.28e-85	0.046
6d	ENSG00000162511	LAPTM5	5.88	1.16e-84	0.046
6d	ENSG00000173175	ADCY5	5.76	6.04e-83	0.046
6d	ENSG00000138944	KIAA1644	5.73	2.28e-84	0.046
6d	ENSG00000127585	FBXL16	5.64	1.51e-84	0.046
6d	ENSG00000141338	ABCA8	5.62	5.34e-84	0.046
6d	ENSG00000187513	GJA4	5.61	2.6e-84	0.046
6d	ENSG00000175229	GAL3ST3	5.57	3.39e-84	0.046
6d	ENSG00000162951	LRRTM1	5.55	1.31e-84	0.046
6d	ENSG00000164161	HHIP	5.49	1.84e-82	0.046
6d	ENSG00000105088	OLFM2	5.36	1.58e-84	0.046
6d	ENSG00000092529	CAPN3	5.21	2.44e-85	0.046
6d	ENSG00000168546	GFRA2	5.18	1.93e-85	0.046
6d	ENSG00000178568	ERBB4	4.79	6.73e-85	0.046
6d	ENSG00000133800	LYVE1	4.77	1.18e-84	0.046
6d	ENSG00000203879	GDI1	4.67	4.04e-85	0.046
6d	ENSG00000151917	BEND6	4.65	1.82e-84	0.046
6d	ENSG00000151552	QDPR	4.63	6.97e-83	0.046
6d	ENSG00000185760	KCNQ5	4.6	6.41e-84	0.046
6d	ENSG00000164588	HCN1	4.59	5.76e-83	0.046
6d	ENSG00000166426	CRABP1	4.57	1.95e-83	0.046
6d	ENSG00000154027	AK5	6.09	7.95e-82	0.048
6d	ENSG00000150656	CNDP1	6.06	3.18e-81	0.048
6d	ENSG00000067840	PDZD4	6	2.64e-81	0.048
6d	ENSG00000157103	SLC6A1	5.96	1.3e-81	0.048
6d	ENSG00000131016	AKAP12	5.75	1.35e-81	0.048
6d	ENSG00000143847	PPFIA4	5.44	8.05e-82	0.048
6d	ENSG00000260230	FRRS1L	4.94	8.16e-82	0.048
6d	ENSG00000113140	SPARC	6.65	3.41e-81	0.049
6d	ENSG00000100505	TRIM9	5.69	5.66e-81	0.049
6d	ENSG00000099864	PALM	5.64	3.61e-81	0.049
6d	ENSG00000089356	FXYD3	-6.04	7.98e-118	0.000
6d	ENSG00000137154	RPS6	-2.05	1.76e-95	0.008
6d	ENSG00000145425	RPS3A	-2.18	2.64e-92	0.023
7	ENSG00000179914	ITLN1	8.77	8.65e-252	0.000
7	ENSG00000158764	ITLN2	7.76	1.88e-196	0.000
7	ENSG00000016490	CLCA1	7.69	1.88e-110	0.000
7	ENSG00000198788	MUC2	7.41	1.77e-152	0.000
7	ENSG00000163515	RETNLB	7.41	8.39e-165	0.000
7	ENSG00000164822	DEFA6	7.38	2.7e-105	0.000
7	ENSG00000130035	GALNT8	7.13	1.84e-263	0.000

Table A.13 Differentially expressed genes in all QC-filtered cells, continued

Cluster	ENSEMBL ID	HGNC symbol	log2FC	p-value	Emp. FDR
7	ENSG00000122711	SPINK4	6.79	2.58e-117	0.000
7	ENSG00000250591	PRSS3P1	6.67	2.08e-133	0.000
7	ENSG00000167748	KLK1	6.64	4.44e-134	0.000
7	ENSG00000106178	CCL24	6.32	5.21e-127	0.000
7	ENSG00000172238	ATOH1	6.25	2.04e-114	0.000
7	ENSG00000184709	LRRC26	6.07	1.08e-106	0.000
7	ENSG00000174562	KLK15	6.04	5.93e-111	0.000
7	ENSG00000169442	CD52	5.17	4.08e-97	0.000
7	ENSG00000142515	KLK3	4.76	4.75e-110	0.000
7	ENSG00000255474	ENSG00000255474	6.15	1.33e-95	0.003
7	ENSG00000197273	GUCA2A	5.62	4.61e-86	0.043

Table A.13: Differentially expressed genes in all QC-filtered cells. The table displays the genes identified as significant (FDR <0.05) in differential expression analysis of clusters identified across all cell types. Genes are ordered by cluster and significance and the log2 fold-change in the cluster compared to all considered single cell libraries is listed for each gene.

A.2.8 Differential expression summary

Cluster	No. DE genes (over-expressed)	No. under-expressed DE genes	Cluster size
1	49	102	516
2	2	19	292
3	20	1	188
4	2	11	295
5	51	2	166
6a	19	0	62
6b	9	0	89
6d	470	3	83
7	18	0	125

Table A.14: Table displays the number of genes significantly (FDR < 0.05) differentially expressed for each single cell BEARscc cluster. The number of differentially expressed genes for each cluster is separated in the number of genes over-expressed in the BEARscc cluster and under-expressed in the BEARscc cluster compared to the rest of the single cell libraries. The BEARscc cluster size is also listed.

A.2.9 BEARscc metrics for clusters from all QC-filtered cells

Cluster	Score	Stability	Promiscuity	Cluster size
1	0.814	0.865	0.051	516
2	0.865	0.893	0.028	292
3	0.972	0.980	0.008	188
4	0.825	0.916	0.091	295
5	0.963	0.980	0.017	166
6	0.586	0.623	0.037	239
7	0.822	0.927	0.105	125

Table A.15: BEARscc metric for the preliminary BEARscc clusters in all-sample analysis.

Cluster	Score	Stability	Promiscuity	Cluster size
1	0.814	0.865	0.051	516
2	0.865	0.893	0.028	292
3	0.972	0.980	0.008	188
4	0.825	0.916	0.091	295

Cluster	Score	Stability	Promiscuity	Cluster size
5	0.963	0.980	0.017	166
6a	0.327	0.980	0.653	62
6b	0.453	0.992	0.539	89
6c	0.132	0.693	0.561	3
6d	0.551	1.00	0.449	83
6e	0.500	1.00	0.500	1
6f	0.480	1.00	0.520	1
7	0.822	0.927	0.105	125

Table A.16: BEARscc metric for the chosen BEARscc clusters in all-sample analysis.

	Specimen 1		**Specimen 2**	
	n	%	*n*	%
MUC2	86	44	107	36
SPINK4	-	-	76	25
ITLN1	36	18	-	-
MUC2+SPINK4	-	-	116	39
MUC2+ITLN1	74	38	-	-
Total	**196**		**299**	

Table A.17: MUC2, ITLN1 and SPINK4 expresssion in two sections from two patients with Barrett's esophagus. The number of cells, n, that scored as positive for expression of each protein by immunofluorescence. The percentage of the total number of scored cells, %, that is positive for the protein listed. See Chapter for details of the antibodies used. All steps and quantification of immunofluorescence were performed by Dr. Michael White.

A.3 List of Software

1. **Samtools** (Li *et al.* 2009)
2. **RSeQC-2.6** (Wang *et al.* 2012)
3. **FastQC** (Andrews 2010)
4. **RStudio** (RStudio Team 2016)
5. **NMF** (Gaujoux & Seoighe 2010)

6. **ggplot2** (Wickham 2009)
7. **Rtsne** (Krijthe 2014; Van Der Maaten 2014; van der Maaten & Hinton 2008)
8. **HTseq** (Anders *et al.* 2015)
9. **STAR V2.4.1** (Dobin *et al.* 2013)
10. **edgeR** (Robinson *et al.* 2009; Chen *et al.* 2014)
11. **DESeq2** (Love *et al.* 2014)
12. **RaceID2** (Grün *et al.* 2016b)
13. **StemID** (Grün *et al.* 2016b)
14. **BackSPIN** (Zeisel *et al.* 2015)
15. **SC3** (Kiselev *et al.* 2016)
16. **Sickle** (Joshi, N A and Fass, J N 2011)
17. **Picard tools** (Wysoker *et al.* 2009)
18. **Kraken** (Wood & Salzberg 2014)
19. **Bowtie2** (Langmead & Salzberg 2012)
20. **Grinder** (Angly *et al.* 2012)

References

1. Ahuja, D., Sáenz-Robles, M. T. & Pipas, J. M. SV40 large T antigen targets multiple cellular pathways to elicit cellular transformation. *Oncogene* **24,** 7729–7745 (Nov. 2005).
2. Aituov, B., Duisembekova, A., Bulenova, A. & Alibek, K. Pathogen-driven gastrointestinal cancers: Time for a change in treatment paradigm? *Infectious Agents and Cancer* **7,** 1–1 (Aug. 2012).
3. Anders, S, Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (Jan. 2015).
4. Andrews, S. *FastQC: a quality control tool for high throughput sequence data* 2010.
5. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research* **40,** e94–e94 (July 2012).
6. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology,* 1–14 (Apr. 2016).
7. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nature Methods* **2,** 731–734 (Oct. 2005).
8. Ballester, V. & Cruz-Correa, M. Endoscopic surveillance of gastrointestinal premalignant lesions. *Current Opinion in Gastroenterology* **30,** 477–483 (Sept. 2014).
9. Barbera, M. *et al.* The human squamous oesophagus has widespread capacity for clonal expansion from cells at diverse stages of differentiation. *Gut* **64,** 11–19 (Dec. 2014).
10. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nature Publishing Group* **15,** 19–33 (Dec. 2013).
11. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449,** 1003–1007 (Oct. 2007).
12. Barker, N. *et al.* Lgr5+ve stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell Stem Cell* **6,** 25–36 (Jan. 2010).
13. Bensley, R. R. The cardiac glands of mammals. *Developmental Dynamics* **2,** 105–156 (Nov. 1902).
14. Booth, C. L. & Thompson, K. S. Barrett's esophagus: A review of diagnostic criteria, clinical surveillance practices and new developments. *Journal of Gastrointestinal Oncology* **3,** 232–242 (Sept. 2012).
15. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10,** 1093–1095 (Sept. 2013).

16. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33,** 155–160 (Jan. 2015).

17. Campbell, J. D. *et al.* The Case for a Pre-Cancer Genome Atlas (PCGA). *Cancer Prevention Research* **9,** 119–124 (Feb. 2016).

18. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541,** 169–175 (Jan. 2017).

19. Cao, Z.-Q. & Guo, X.-L. The role of galectin-4 in physiology and diseases. *Protein & Cell* **7,** 314–324 (Mar. 2016).

20. Carulli, A. J., Samuelson, L. C. & Schnell, S. Unraveling intestinal stem cell behavior with models of crypt dynamics. *Integrative Biology* **6,** 243–257 (Mar. 2014).

21. Chen, Y *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29,** 266–267 (Jan. 2013).

22. Chen, Y., McCarthy, D. J., Robinson, M. & Smyth, G. K. edgeR: differential expression analysis of digital gene expression data. *Bioconductor Manual,* 1–79 (Sept. 2014).

23. CRUK. *Cancer Research UK: Oesophageal cancer statistics* http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#heading-Six. [Online; accessed 22 February 2017]. 2014.

24. Cunningham, D., Okines, A. F. C. & Ashley, S. Capecitabine and oxaliplatin for advanced esophagogastric cancer. *The New England Journal of Medicine* **362,** 858–859 (Mar. 2010).

25. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 7285–7290 (June 2015).

26. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (Jan. 2013).

27. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nature Publishing Group* **14,** 20–32 (Oct. 2015).

28. Drewitz, D. J. & Sampliner, R. E. The incidence of adenocarcinoma in Barrett's esophagus: a prospective study of 170 patients followed 4.8 years. *American Journal of Gastroenterology* (1997).

29. Ennen, M *et al.* Single-cell gene expression signatures reveal melanoma cell heterogeneity. *Oncogene* **34,** 3251–3263 (Aug. 2014).

30. Faridi, R., Zahra, A., Khan, K. & Idrees, M. Oncogenic potential of Human Papillomavirus (HPV) and its relation with cervical cancer. *Virology Journal* **8,** 269 (June 2011).

31. Forster, M. *et al.* Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Scientific Reports* **5,** 11534 (2015).

32. Garman, K. S., Orlando, R. C. & Chen, X. Review: Experimental models for Barrett's esophagus and esophageal adenocarcinoma. *AJP: Gastrointestinal and Liver Physiology* **302,** G1231–G1243 (June 2012).
33. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11** (2010).
34. Gilbert, E. W., Luna, R. A., Harrison, V. L. & Hunter, J. G. Barrett's Esophagus: A Review of the Literature. *Journal of Gastrointestinal Surgery* **15,** 708–718 (Apr. 2011).
35. Gilbert, J. Establishment of permanent cell lines by Epstein-Barr virus transformation. *Current protocols in human genetics* **Appendix 3,** Appendix 3H–A.3H.5 (May 2001).
36. Grün, D, Muraro, M. J., Boisset, J. C. & Wiebrands, K. De novo prediction of stem cell identity using single-cell transcriptome data - Supplemental. *Cell Stem Cell* (2016).
37. Grün, D. & van Oudenaarden, A. Design and analysis of single-cell sequencing experiments. *Cell* **163,** 799–810 (Nov. 2015).
38. Grün, D. *et al.* De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19,** 266–277 (Aug. 2016).
39. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (Aug. 2015).
40. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11,** 637–640 (Apr. 2014).
41. Gulley, M. L. Genomic assays for Epstein-Barr virus-positive gastric adenocarcinoma. *Experimental & Molecular Medicine* **47,** e134 (2015).
42. Heath, J. P. Epithelial cell migration in the intestine. *Cell biology international* **20,** 139–146 (1996).
43. Hemme, C. L. *et al.* Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. *Frontiers in Microbiology* **6,** 578–512 (Oct. 2015).
44. Hicks, S. C., Teng, M & Irizarry, R. A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* (2015).
45. Hoffmann, W. Regeneration of the gastric mucosa and its glands from stem cells. *Current Medicinal Chemistry* (2008).
46. Hoffmann, W. Current Status on Stem Cells and Cancers of the Gastric Epithelium. *International Journal of Molecular Sciences* **16,** 19153–19169 (Aug. 2015).
47. Hoffmann, W. Self-renewal of the gastric epithelium from stem and progenitor cells. *Frontiers in bioscience (Scholar edition)* **5,** 720–731 (Jan. 2013).
48. Holmberg, D., Ness-Jensen, E., Mattsson, F., El-Serag, H. B. & Lagergren, J. Risk of oesophageal adenocarcinoma in individuals with Barrett's oesophagus. *European Journal of Cancer* **75,** 41–46 (Apr. 2017).
49. Hu, Y. *et al.* The Pathogenesis of Barrett's Esophagus: Secondary Bile Acids Upregulate Intestinal Differentiation Factor CDX2 Expression in Esophageal Cells. *Journal of Gastrointestinal Surgery* **11,** 827–834 (Apr. 2007).

50. Hudnall, S. D. *Viruses and Human Cancer* (Springer, June 2014).

51. Huse, J. T. & Aldape, K. CMV and glioma-are we there yet? *Neuro-Oncology* **16,** 1433–1434 (Oct. 2014).

52. Hutchinson, L, Stenstrom, B & Chen, D. Human Barrett's adenocarcinoma of the esophagus, associated myofibroblasts, and endothelium can arise from bone marrow-derived cells after allogeneic stem cell transplant. *Stem cells and Development* (2010).

53. Hvid-Jensen, F, Pedersen, L & Drewes, A. M. Incidence of Adenocarcinoma among Patients with Barrett's Esophagus. *New England Journal of Medicine* **365,** 1375–1383 (2011).

54. Jang, B. G., Lee, B. L. & Kim, W.-H. Intestinal Stem Cell Markers in the Intestinal Metaplasia of Stomach and Barrett's Esophagus. *PLoS ONE* **10,** e0127300–13 (May 2015).

55. Jansen, M. & Wright, N. A. Distal Esophageal Adenocarcinoma and Gastric Adenocarcinoma: Time for a Shared Research Agenda. *Advances in experimental medicine and biology* **908,** 1–8 (2016).

56. Jiang, L *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Research* **21,** 1543–1551 (Sept. 2011).

57. Joshi, N A and Fass, J N. *Sickle: A sliding window, adaptive, quality-based, trimming tool for FastQ files (Version 1.33) [Software]* Available at https://github.com/najoshi/sickle. 2011.

58. Kazumori, H. Bile acids directly augment caudal related homeobox gene Cdx2 expression in oesophageal keratinocytes in Barrett's epithelium. *Gut* **55,** 16–25 (Jan. 2006).

59. KC, K., Rothenberg, M. E. & Sherrill, J. D. In Vitro Model for Studying Esophageal Epithelial Differentiation and Allergic Inflammatory Responses Identifies Keratin Involvement in Eosinophilic Esophagitis. *PLoS ONE* **10,** e0127755–18 (June 2015).

60. Kim, J. K., Kolodziejczyk, A. A., Illicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications* **6,** 8687–8688 (Oct. 2015).

61. Kim, M. J., MSc, G. M. S. & Yang, H. Effect of human papillomavirus vaccination on cervical cancer screening in Alberta. *CMAJ* (2016).

62. Kim, T.-H. & Shivdasani, R. A. Stomach development, stem cells and disease. *Development* **143,** 554–565 (Feb. 2016).

63. Kim, T.-H. *et al.* Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity. *Nature* **506,** 511–515 (Jan. 2014).

64. Kiselev, V. Y., Kirschner, K, Schaub, M. T. & Andrews, T. SC3-consensus clustering of single-cell RNA-Seq data. *bioRxiv* (2016).

65. Kostic, A. D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnology* **29,** 393–396 (May 2011).

66. Krijthe, J. H. *Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation* 2014. <https://github.com/jkrijthe/Rtsne>.

67. Kutlubay, Z., Engin, B., Zara, T. & Tüzün, Y. Anogenital malignancies and premalignancies: facts and controversies. *Clinics in Dermatology* **31,** 362–373 (July 2013).

68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (Mar. 2012).

69. Lavery, D. L. *et al.* The stem cell organisation, and the proliferative and gene expression profile of Barrett's epithelium, replicates pyloric-type gastric glands. *Gut* **63,** 1854–1863 (Oct. 2014).

70. Li, H *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (Aug. 2009).

71. Liu, Y. *et al.* Genome-wide analysis of Epstein-Barr virus (EBV) isolated from EBV-associated gastric carcinoma (EBVaGC). *Oncotarget* (Dec. 2015).

72. Long, J. D. & Orlando, R. C. Esophageal submucosal glands: structure and function. *The American Journal of Gastroenterology* **94,** 2818–2824 (Oct. 1999).

73. Lörinc, E. & Öberg, S. Hyperplasia of the submucosal glands of the columnar-lined oesophagus. *Histopathology* **66,** 726–731 (Jan. 2015).

74. Lörinc, E., Mellblom, L. & Öberg, S. The immunophenotypic relationship between the submucosal gland unit, columnar metaplasia and squamous islands in the columnar-lined oesophagus. *Histopathology* **67,** 792–798 (June 2015).

75. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 31–21 (Dec. 2014).

76. Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Reports* **7,** 1130–1142 (May 2014).

77. McDonald, S. A. C., Lavery, D., Wright, N. A. & Jansen, M. Barrett oesophagus: lessons on its origins from the lesion itself. *Nature Reviews Gastroenterology & Hepatology* **12,** 50–60 (Nov. 2014).

78. McDonald, S. A. C., Graham, T. A., Lavery, D. L., Wright, N. A. & Jansen, M. The Barrett's Gland in Phenotype Space. *Cellular and Molecular Gastroenterology and Hepatology* **1,** 41–54 (Jan. 2015).

79. National Library of Medicine US, National Center for Biotechnology Information. *The NCBI handbook [Internet]* Access date: August 7th 2014. 2002.

80. Ohashi, S. *et al.* NOTCH1 and NOTCH3 Coordinate Esophageal Squamous Differentiation Through a CSL-Dependent Transcriptional Network. *Gastroenterology* **139,** 2113–2123 (Dec. 2010).

81. Parsons, T. E. *Foregut microbiome in development of esophageal adenocarcinoma* (Masters Abstracts International, 2011).

82. PathologyOutlines.com. *PathologyOutlines.com: Esophagus - General Histology* http://www.pathologyoutlines.com/topic/esophagusnormalhistology.html. Accessed: 16 April 2016. 2013.

83. Peek, R. M. & Blaser, M. J. Helicobacter pylori and gastrointestinal tract adnocarcinomas. *Nature Reviews Cancer* **2,** 28–37 (Jan. 2002).

84. Pellegrinet, L. *et al.* Dll1- and dll4-mediated notch signaling are required for homeostasis of intestinal stem cells. *Gastroenterology* **140,** 1230–1240.e1–7 (Apr. 2011).

85. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9,** 171–181 (Jan. 2014).

86. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10,** 1096–1098 (Sept. 2013).

87. Quante, M. *et al.* Bile Acid and Inflammation Activate Gastric Cardia Stem Cells in a Mouse Model of Barrett-Like Metaplasia. *Cancer Cell* **21,** 36–51 (Jan. 2012).

88. Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* (1971).

89. Reddout, N. *et al.* High risk HPV types 18 and 16 are potent modulators of oral squamous cell carcinoma phenotypes in vitro. *Infectious Agents and Cancer* **2,** 21–11 (2007).

90. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (Dec. 2009).

91. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11,** 396–398 (Apr. 2014).

92. RStudio Team. *RStudio: Integrated Development for R.* 2016. <http://www.rstudio.com/>.

93. Salo, J. A., Kivilaakso, E. O. & Kiviluoto, T. A. Cytokeratin profile suggests metaplastic epithelial transformation in Barrett's oesophagus. *Annals of Medicine* (1996).

94. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12,** 87 (2014).

95. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459,** 262–265 (May 2009).

96. Savarino, E *et al.* The natural history of gastro-esophageal reflux disease: a comprehensive review. *Diseases of the Esophagus* **30,** 1–9 (Nov. 2016).

97. Sharma, P. Are screening and surveillance for Barrett's oesophagus really worthwhile? *Gut* **54,** i27–i32 (Mar. 2005).

98. Shin, J *et al.* The intestinal epithelial cell differentiation marker intestinal alkaline phosphatase (ALPi) is selectively induced by histone deacetylase inhibitors (HDACi) in colon cancer cells in a Kruppel-like factor 5 (KLF5)-dependent manner. *The Journal of Biological Chemistry* **289,** 25306–25316 (Sept. 2014).

99. Söderberg-Nauclér, C. Cytomegalovirus in human brain tumors: Role in pathogenesis and potential treatment options. *World Journal of Experimental Medicine* **5,** 1–11 (2015).

100. Söderberg-Nauclér, C. & Johnsen, J. I. Cytomegalovirus infection in brain tumors. *OncoImmunology* **1,** 739–740 (Oct. 2014).

101. Solaymani-Dodaran, M. Risk of oesophageal cancer in Barrett's oesophagus and gastro-oesophageal reflux. *Gut* **53,** 1070–1074 (Aug. 2004).

102. Stachler, M. D. *et al.* Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nature Genetics* **47,** 1047–1055 (July 2015).

103. Takubo, K. *et al.* The normal anatomy around the oesophagogastric junction: A histopathologic view and its correlation with endoscopy. *Best Practice & Research Clinical Gastroenterology* **22,** 569–583 (Aug. 2008).

104. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature,* 1–8 (July 2014).

105. Theron, B. T. *et al.* The risk of oesophageal adenocarcinoma in a prospectively recruited Barrett's oesophagus cohort. *United European Gastroenterology Journal* **4,** 754–761 (Dec. 2016).

106. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (Apr. 2016).

107. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology (Pozna)* **19,** A68–77 (2015).

108. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology* **11,** e1004333–18 (June 2015).

109. Van Es, J. H. *et al.* Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. *Nature Cell Biology* **14,** 1099–1104 (Sept. 2012).

110. Van Nieuwenhove, Y, Destordeur, H & Willems, G. Spatial distribution and cell kinetics of the glands in the human esophageal mucosa. *European Journal of Morphology* **39,** 163–168 (July 2001).

111. Van der Flier, L. G., Haegebarth, A, Stange, D. E., van de Wetering, M & Clevers, H. OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. *YGAST* **137,** 15–17 (July 2009).

112. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (Nov. 2008).

113. Van der Veen, A. H., Dees, J, Blankensteijn, J. D. & Van Blankenstein, M. Adenocarcinoma in Barrett's oesophagus: an overrated risk. *Gut* **30,** 14–18 (Jan. 1989).

114. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* (2014).

115. Vennalaganti, P. *et al.* Discordance Among Pathologists in the United States and Europe in Diagnosis of Low-Grade Dysplasia for Patients With Barrett's Esophagus. *Gastroenterology* **152,** 564–570.e4 (Feb. 2017).

116. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* **34,** 1145–1160 (Nov. 2016).

117. Wang, J. *et al.* Differential gene expression in normal esophagus and Barrett's esophagus. *Journal of Gastroenterology* **44,** 897–911 (May 2009).

118. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics* **46,** 573–582 (May 2014).

119. Wang, L, Wang, S & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28,** 2184–2185 (Aug. 2012).

120. Wang, X. *et al.* Residual Embryonic Cells as Precursors of a Barrett's-like Metaplasia. *Cell* **145,** 1023–1035 (June 2011).

121. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics* **45,** 1113–1120 (Oct. 2013).

122. Wick, W & Platten, M. CMV infection and glioma, a highly controversial concept struggling in the clinical arena. *Neuro-Oncology* **16,** 332–333 (Feb. 2014).

123. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2009).

124. Wood, D & Salzberg, S. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* (2014).

125. Wysoker, A., Tibbetts, K. & Finnell, T. *Picard tools version 1.90* 2009. <http://picard.sourceforge.net>.

126. Yamamoto, Y. *et al.* Mutational spectrum of Barrett’s stem cells suggests paths to initiation of a precancerous lesion. *Nature Communications* **7,** 1–10 (Jan. 2016).

127. Ye, W *et al.* Helicobacter pylori Infection and Gastric Atrophy: Risk of Adenocarcinoma and Squamous-Cell Carcinoma of the Esophagus and Adenocarcinoma of the Gastric Cardia. *Journal of the National Cancer Institute* **96,** 388–396 (Mar. 2004).

128. Zagari, R. M. *et al.* Prevalence of upper gastrointestinal endoscopic findings in the community: A systematic review of studies in unselected samples of subjects. *Journal of Gastroenterology and Hepatology* **31,** 1527–1538 (Sept. 2016).

129. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).