Computational analysis of intercellular communication in APC-driven colorectal cancers with varying KRas mutational status

by

Stephen C Van Nostrand

B.A., Rutgers University (2017)

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Master's of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author . . . .
Department of Biological Engineering
August 7, 2020

Certified by . . . .
Douglas A Lauffenburger
Ford Professor of Biological Engineering
Thesis Supervisor

Accepted by . . . .
Katharina Ribbeck
Chairman, Department Committee on Graduate Theses

Computational analysis of intercellular communication in APC-driven colorectal cancers with varying KRas mutational status

by

Stephen C Van Nostrand

Submitted to the Department of Biological Engineering
on August 7, 2020, in partial fulfillment of the
requirements for the degree of
Master's of Science

Abstract

Three common KRas mutants were compared with wild type KRas in a mouse model of APC-driven colorectal cancer to understand differences in cell-cell communication. Using single-cell RNA sequencing and a handful of novel computational methods, a set of nine highest priority ligands expressed by non-immune cells that differ statistically between G13D-mutant KRas and the others was identified for further study. This set contains two ligands that have been previously recognized as important in this context, as well as novel ligands and some with poorly understood relevance to the clinic. While no secondary validation of how these ligands could be affecting clinical outcomes was performed here, the simplicity of interpretation of the computational methods demonstrated begs for further study, particularly of the effects of changes in these ligands in vivo. Follow-up studies will be undertaken at the Dana Farber Cancer Institute to continue fleshing out our understanding of how molecular differences in KRas can lead to differences in tumor composition as well as distinct prognoses.

Thesis Supervisor: Douglas A Lauffenburger
Title: Ford Professor of Biological Engineering

Acknowledgments

Many thanks to my advisor, Douglas A Lauffenburger, without whom neither this project or I would have gotten anywhere. From when I met him as BE Department Head during Interview Weekend in 2017, Doug always showed how much he cared not only for his students in the broadest sense but for each person individually as well. He's provided me a lot of much-needed support, both academic and otherwise.

I also relied on the support of collaborators, particularly those on the project presented here – Ken S Lau and Kevin M Haigis – but also those of previous projects – chiefly among them Elizabeth A Proctor and Barbara S Nikolajczyk. These researchers as well as others in their labs generated data for me to study, discussed interesting questions, and explained the relevant background. In particular, I would also like to thank Yi-Jang Lin and Moon Hee Yang (Haigis), Bob Chen (Lau), and Madhur Agrawal (Nikolajczyk).

Another layer of support also came from the members of the Griffenburger lab. Of particular note, Brian A Joughin met with me innumerable times to help me with my various projects, as well as championed data meeting to foster interactions and practice communicating. My other mentors in the lab, particularly Hsinhwa and Alina, as well as all our labmates were also indispensable for their help and input.

I couldn't have made it here without every one of my friends, both nearby or geographically far. Amanda, Sarah, Caitlyn, and Nick; the Mudd Squad; the BE department and my cohort in particular; those at MIT and those not – to list why I'm thankful for each would take too many pages to reasonably list.

Patrick, you've been alongside me for this whole process and you've contributed so much to the completion of this work in so many invisible ways. I've relied on both your brain and your heart more than I'm keen to admit.

And finally, to Mom and Dad – as well as the rest of my family: Peter, Grandma and Grandpa, Aunt/Anut Lynn, Andrew, Sam and Bash, Grammy and Grampie – I give my utmost thanks and love. You've all made sure I take care of myself and don't feel alone even when I'm overwhelmed. Your advice, caring, and love keep me sane.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 KRas in Colorectal Cancer

Colorectal cancer (CRC) is the fourth-most common cancer worldwide, as well as the second-most deadly, and its incidence is continuing to rise, particularly in non-Western countries. The five-year survival rate in the United States across all stages is only 64.5%.[30] Nearly all ($\sim$ 96%) of these cancers originate as polyps [30] – non-invasive growths of the mucus-producing cells in the walls of the large intestine – often with mutations in the tumor suppressor p53 or a common oncogene such as APC or KRas.[32] Risk factors that increase the likelihood of developing CRC are age, inflammatory bowel disease (IBD), obesity, and diabetes.[30]

The current standard of care is surgical resection for early stages, adjuvant chemotherapy for later stages, and cytotoxic chemotherapy for metastatic CRC.[28] Anti-EGFR antibodies are often used in combination with chemotherapy, however they are only used in patients with wild type KRas as KRas mutational status is highly predictive of resistance to EGFR inhibition.[23]

KRas is an oncogene mutated in approximately 40% of patients with CRC, and is the most commonly mutated oncogene across cancer as a whole – lying downstream of EGFR and upstream of the MAPK/ERK, PI3K, and NFκB pathways (among others) leading to increased survival and proliferation.[31, 10, 21] Like other small GTPases, KRas is active when bound to a high-energy GTP molecule and deactivates upon hydrolysis to the lower-energy GDP, requiring the protein to exchange its bound GDP for a new GTP molecule to reactivate. Mutations in KRas and its associated pathways correlate with resistance to EGFR inhibition, leading to poor prognosis for KRas-mutant cancers, which account for more than 30% of all cancer patients.[28, 25]

Studies across many laboratories and institutions have found differences between the different mutants of KRas. The most common site for mutation is G12, found in approximately 80% of CRC patients with a KRas mutation,[27] however other sites, such as G13 and A146, are reasonably common as well. As a GTPase, many of these mutations occur in the nucleotide binding and hydrolysis region – G12 mutations slow intrinsic and extrinsically-driven hydrolysis, A146 mutations drastically raise nucleotide exchange rates, and G13 mutations affect both processes to a lesser degree. These molecular effects also translate into differences at the epidemiological level – G12 mutations are significantly correlated with worse prognosis and lower survival rates, while A146 mutations have improved survival compared to others. These differences are related to differences in therapeutic responses such as inhibition of EGFR with e.g. cetuximab – to which G12-mutant CRCs are resistant while G13 mutants' responses are between those of G12-mutant and KRas-wild type tumors.[10] While KRas was originally divided only by whether it was mutated or not,[23] this view has been superseded by a more personalized view based on which mutation is present in a patient's tumor that can hopefully better guide treatment.[10]

Downstream of KRas along multiple signaling pathways lie intercellular signals – proteins that are secreted or displayed on the plasma membrane to communicate with other cells. Cancer-associated fibroblasts and both myeloid- and lymphoid-derived immune cells are targets for some of these signals, through various cytokines, chemokines, and growth factors.[21] These signals also affect other malignant cells to promote many malignant cellular phenotypes.[1] Tumor cell signaling can have large effects on their microenvironment: GM-CSF signaling recruits myeloid-derived suppressor cells to minimize immune response, the immunotherapy target PD-L1 selectively excludes B and T cells, and IL-6 reprograms the stroma to increase angiogenesis.[21, 1]

1.2 scRNAseq & Cell-cell communication methods

One technology that has been increasingly used to study intercellular communication, including in CRC, is single-cell RNA sequencing (scRNAseq). Due to its high-dimensional nature, scRNAseq takes a broad snapshot of what individual cells are doing within a sample, allowing for inference of both cell types and clonal heterogeneity of expression. Many previous studies have used this to their advantage to quantify signaling and signal strength between different cell types.[4, 5, 11, 19, 36]

Broadly, scRNAseq works by isolating single cells in suspension, encapsulating them (usually in droplets or wells), lysing the cells, and converting their mRNA to a complementary DNA template with unique barcodes to identify each cell and transcript, followed by more generic DNA amplification and sequencing library preparation. This method allows for counting reads from the same transcript only once, and grouping these counts based on the transcripts' cell of origin. With this information, expression is quantified on the single-cell, single-molecule scale, generating huge amounts of data to be handled computationally.

Using the expression of genes whose protein products are known markers of specific cell types, the barcodes corresponding to cells of the same type can be grouped to generate distributions and draw general conclusions about each cell type individually. Furthermore, using lists of known ligand-receptor pairs (such as provided in [29]), a number of computational methods have been developed to infer protein-level interactions between cells based on these RNA-level measurements.

Seemingly the most common way of evaluating cell-cell communication in a high-throughput fashion is to take some combination of the RNA expression of both the ligand and its cognate receptor. Some take a simple "score" for the interaction as the product of the mean ligand and receptor expressions – optionally grouping either by cell type for a more nuanced look at what cells are communicating.[18, 19, 36] Others have used the mean of both genes,[5] more complicated combinations such as the L^2-norm of the z-score of each gene across all cell types,[11], or even complex graph-based analysis for relative significance.[4]

One thing that all of these methods have in common is that they look at the strength of both the receptor and the ligand when considering which interactions are likely occuring in situ. However, this ignores situations where, for example, the signaling network only requires low absolute amounts of activated receptors to produce a large response to the ligand. Additionally, when comparing interactions between different samples, they generally treat a two-fold difference in the ligand as equivalent to a two-fold difference in the receptor – which is not always the case when, for example, the receptors are already nearing saturation with the lower amount of ligand and the equilibrium leads to a sublinear response. Here, we instead look only at changes in the amount of ligand as long as the receptor is expressed at a level capable of receiving signal. Therefore, the differences in ligand expression are treated as the only value changing – which is particularly useful with an assumption that signals of interest are unidirectional.

1.3 Goals of this work

In the current study, we aim to use single-cell RNA sequencing (scRNAseq) to evaluate differences in cell-cell communication across genetically-engineered mouse models of colorectal cancer (CRC) with different KRas mutations (using APC deletion specifically within the gastrointestinal tract to initiate malignancy). While KRas is known to drive differences in signaling, and different KRas mutations are known to have differing clinical consequences for patients with CRC, this work seems to be the first to examine the signaling mechanisms that could lead from molecular differences to clinical outcomes.

Here, intercellular signaling is computationally inferred downstream of wild type KRas and three of its most common mutations in CRC (G12D, G13D, and A146T), with specific focus placed on findings that tumors with G13D-mutant KRas have increased immune cell infiltration. A prioritized list of ligands that are significantly changed in non-immune cells from the G13D model is presented for experimental follow-ups (such as ligand overexpression or blocking ligand-receptor interaction) to confirm phenotypic and clinical relevance. These are thought of as signals produced by the tumor cells themselves that could be causing mutation-specific differences in their microenvironment. The hope is that eventually, the ligands of interest here could expand both the biological understanding of specific KRas mutations as well as inform therapeutic design and use that is more tailored to smaller subsets of patients.

Chapter 2

Methods

2.1 Animal protocols

Three mice from each of the four KRas genotypes under study were used, one of which from each model was additionally enriched for CD45+ cells to ensure full interrogation of the immune compartment (see Section 2.3). All of the mice were Apc$^{fl/fl}$ with inducible Villin-CreER, with mutant KRas alleles containing a lox-stop-lox motif. This ensures that APC deletion and KRas mutation occur specifically in the gastrointestonal tract (the site of nearly all Villin expression [35]) when Cre recombinase expression is induced.

Mice were orthotopically injected with 4-hydroxytamoxifen to induce tumor formation, and tumor growth was monitored each week by colonoscope. KRas mutation had no effect on tumor initiation, but led to faster growth compared to wild type KRas. When the colonic lumen was 50% obstructed due to tumor growth, the tumors were harvested and non-malignant epithelium trimmed away from the palpable tumor mass.

2.2 RNA extraction and sequencing

Tumor dissociation into single cells for library preparation was done as a two-step process, as described previously.[22, 13] In the first step, harvested tumors were washed in ice-cold PBS and digested in 2 mg/mL collagenase type II in DMEM at 37° for 1 hr or until fragments had dispersed. After washing again with ice-cold PBS, the suspension was filtered to isolate epithelial crypts larger than 40 µm and smaller than 100 µm. In the second step, the suspension was chelated with a 3 mM EDTA and 1 mM DTT buffer at 4° for 45 min followed by 30 s of shaking. Crypts were then digested again with 2 mg/mL Collagenase type II and 2.5 mg/mL DNAse at 37° for 20 min followed by mechanical disruption with a 27.5-gauge needle. The resulting single cell suspension was washed twice with ice-cold PBS and enriched for live cells using the MACS dead cell removal kit (magnetic-activated cell sorting; from Miltenyi Biotec). Final suspensions were prepared at $1.5 * 10^5$ cells/mL, which were spiked with 18 µL Optiprep per 100 µL suspension to maintain cell viability.[22, 13]

Encapsulation was performed using the inDrop platform (from 1CellBio) as described previously.[17] Using the CEL-Seq workflow, the inDrop protocol begins with reverse transcription followed by exonuclease I (ExoI) digestion and solid-phase reversible immobilization purification (SPRIP). Then single-strand synthesis and another round of SPRIP generates a pure cDNA template for in vitro transcription linear amplification using T7 polymerase (followed by SPRIP) to generate many copies of RNA for each original molecule. Finally, RNA fragmentation (followed by SPRIPR), library primer ligation, final reverse transcription, and enrichment PCR yield a large DNA library with the proper sequencing primers.

Resulting libraries were sequenced on a BGI nanoball system, splitting each sample across four lanes for deeper sequencing and therefore more accurate quantification.

2.3 CD45 enrichment

CD45 enrichment of one sample per KRas genotype was done using the MACS kit as per the manufacturer's directions except replacing the 0.5% BSA with 0.05% BSA and using large (LS) columns. Briefly, dissociated cells were resuspended in buffer containing magnetic microbeads coated with anti-CD45 antibodies. Then a magnet was used to hold the captured cells in a filter column while the unbound cells were washed

Table 2.1: Description of the samples

Sample	KRas	CD45 Enriched?	Mouse number	Sac Date	# barcodes detected	# likely cells
108RD	WT	Yes	2602	08 Nov 2018	8532	2516
144R		No			18960	4631
178D	WT	No	2625	09 Nov 2018	17103	3246
200	WT	No	2700	08 Nov 2018	7033	1975
110RD	G12D	Yes	2622	09 Nov 2018	10837	1856
146R		No			10975	2524
175D	G12D	No	2627	07 Nov 2018	9382	2379
201	G12D	No	2626	09 Nov 2018	7106	1687
297	G13D	Yes	5764	20 Mar 2019	7622	1930
199R		No			16952	4425
177D	G13D	No	5649	22 Feb 2019	8691	1924
203	G13D	No	5763	20 Mar 2019	7443	1930
109RD	A146T	Yes	2607	08 Nov 2018	4009	1115
145R		No			16523	3021
176D	A146T	No	2603	07 Nov 2018	17461	4254
202	A146T	No	2610	07 Nov 2018	12109	2094

through with buffer. Finally, the bead-bound cells were eluted from the column using buffer after removal from the magnet. These enriched samples then continued on with library preparation and sequencing as above, alongside an unenriched fraction from the same tumor dissociation.

2.4 Raw data processing

Demultiplexing of cell barcodes was done using dropTag for each lane using the default configuration file for inDrop v1/2, also reporting base call quality scores in the UMI (unique molecular identification; the barcode given to each transcript) using the `-s` flag.[26] For each lane, dropTag takes in the paired-end reads – one being the cell barcode and UMI, while the other is a part of the gene – and produces a single `.fastq` file of biological (i.e. gene) reads annotated with sample, cell barcode, and UMI metadata.

Single-end alignment of the biological reads against the mouse mm10 (GRCm38) transcriptome [3] was then performed using the TopHat2 aligner [15] and the corresponding BowTie2 index [20], prepared with the annotation file (`.gtf` format) available on Ensembl [38], reporting only the best match for each read. This step also combined the four lanes of each sample, yielding a single alignment file.

The resulting alignments were then counted using dropEst, given the base call qualities from dropTag and using the barcode correction method via the `-m` flag.[26] For each sample, dropEst merges cell barcodes with errors into a canonical list provided for the library preparation procedure, and merges UMI sequences that are close to each other using an estimate of the likelihood that they originally derived from the same molecule in the cell and were split due to an error during amplification or sequencing. dropEst returns a matrix of every unique barcode (less than the number of barcodes found in sequencing due to merging of errors) by every gene that was captured in the biological reads in any cell of that sample (from alignment), where the value is the estimated number of truly unique UMIs (i.e. corrected for likely errors).

This "count matrix," as it is known, is assumed to be a faithful count of the number of unique transcripts for each gene that was captured during library preparation. Note that while this is likely a good assumption for highly expressed genes – where missing a molecule or two of mRNA is only a small percentage difference in expression – the lower the biological expression of the gene, the higher the chance of missing a significant portion of the mRNA molecules for that gene, leading to a preponderance of zeros referred to as "drop outs" (i.e. genes that aren't measured despite being expressed in the cell). Therefore, there is systemic bias in scRNAseq data that makes lowly expressed genes seem even less expressed than they are in reality.

Steps 1–3 (see Table 2.2) were done on the `luria` cluster at MIT that runs CentOS, using 16 Intel Xeon E5-2650 v2 (2.60 GHz) or E5620 (2.40 GHz) CPUs and up to 128 GB of memory, except for Step 1 that used only 8 cores with up to 96 GB, 128 GB, or 192 GB of memory.

These steps used R v3.5.1, BamTools v2.5.1, dropEst v0.8.5, TopHat v2.1.1, BowTie v2.3.5.1, OpenBLAS v0.2.19, and Pandoc v2.7.3. All further computation was performed on an ASUS Q534UXK laptop with a Intel Core i7-7500U CPU (2.70 GHz, 2904 MHz) running Windows 10 Home. 64-bit Python v3.7.1 and its modules were provided by Anaconda3, including NumPy v1.16.2, pandas v0.23.4, Matplotlib v3.0.2, SciPy v1.1.0, and scikit-learn v0.20.1 and run using a Jupyter notebook v5.7.4 with an iPython v7.2.0 kernel.

2.5 Preprocessing, Filtering, & Classification

Count matrices were loaded into a python environment using the rpy2 module v2.9.1 from the `.rds` files output by dropEst. The sparse (barcode- i.e. column-oriented) `dgCMatrix` was converted to a dense array using a custom function. The data were then loaded into custom objects for further processing and eventual analysis.

Barcodes were filtered in each sample based on total counts as described in [12]. Briefly, given a count matrix f: 1, barcodes were arranged by decreasing total counts; 2, the cumulative sum was calculated; and 3, the secant line was drawn from $(0, 0)$ to $\left(N, \sum_{ij} f_{ij}\right)$. The distance between the cumulative sum and its secant was then used as a measure of the contribution of each barcode to the total counts across the sample. Cells with more total counts than the 70th percentile of this distance after its maximum were kept.

The data were then normalized using a minor modification of the gene frequency / inverse cell frequency (GF-ICF) method, itself previously adapted from the term frequency / inverse document frequency (TF-IDF) score used in text mining to rank doc-

Table 2.2: Data processing time

#	Step	Avg Time $(hh : mm)$	Runs	Total Time $(hh : mm)$
1	dropTag demultiplexing	00 : 12	64	12 : 31
2a	TopHat2 genome preparation	00 : 19	1	00 : 19
2b	TopHat2 alignment	02 : 44	16	43 : 55
3	dropEst count estimation	00 : 30	16	07 : 54
	Total			64 : 40

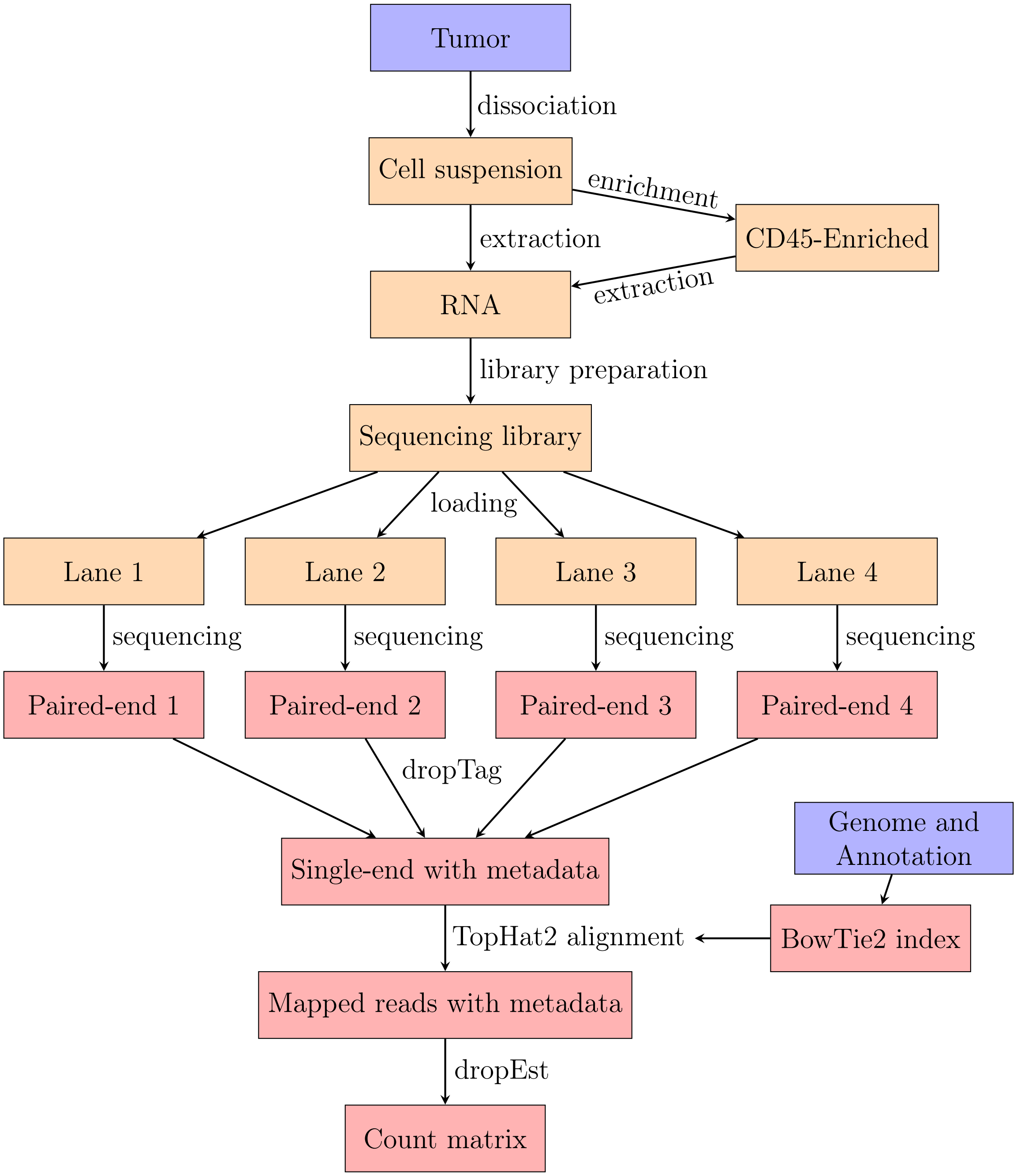

Figure 2-1: Flow diagram of processing steps

Purple are inputs, orange is a physical sample, red is a data file uments based on the frequency of a given word across a large set of text documents.[8] Briefly, given a count matrix f, the gene frequency of gene i in cell j is

$$gf_{ij} = \frac{f_{ij}}{\sum_i f_{ij}},$$

or the fraction of UMIs in that cell from the given gene and is equivalent to $CPM/10^6$. Note that $\sum_i f_{ij} = 0$ iff no genes were detected for cell j, and in those cases all gf_{ij} are taken to be 0.

The inverse cell frequency of gene i is

$$icf_i = \begin{cases} \log\left(\frac{N}{n_i} + 1\right) & n_i \neq 0 \\ 0 & n_i = 0, \end{cases}$$

where $0 \leq n_i \leq N$ is the number of cells in which at least one UMI from the given gene was seen. The base of the logarithm is arbitrary, as a change of base is equivalent to constant multiplicative scaling. This inverse frequency is high for genes that are captured in few cells and low for genes captured in many cells, thereby providing gene-wise weighting that exaggerates uncommon genes and minimizes common genes.

The value taken as the "expression" of gene i in cell j is then $gf_{ij} * icf_i$. Unlike in [8], L^2-normalization was not performed as that removes the dependence on the total counts in a cell (see Appendix A). Additionally, for better handling of the extrema of n_i, the above definition of icf_i has been extended from that given in [8]: $icf_i = log\left(\frac{N}{n_i} + 1\right)$. The modified form above returns 0 for $n_i = 0$ instead of ∞, which allows undetected genes to be dropped before or after normalization. As this normalization scheme is dependent on both the counts within each cell as well as the detection of a gene across all cells, GF-ICF normalization needs to repeated from the raw counts when changing which cells are included.

The first round of filtering was done for each sample separately. The sample was clustered using the Leiden algorithm as implemented in SCANPY.[34, 37] Each cell is treated as a node of a graph and edges are drawn between cells with non-zero membership in the fuzzy simplicial sets calculated as an intermediate in UMAP after

PCA transformation.[37, 24] Each node is assigned to its own cluster, and clusters are merged greedily if they improve (i.e. lower) the Constant Potts Model (CPM; note that this is not the "counts per millions" used elsewhere) objective function $\mathcal{H}$ of the partitioning $\mathcal{P}$, i.e.

$$\mathcal{H}(\mathcal{P}) = \sum_{c \in \mathcal{P}} \left[e_c - \gamma \binom{n_c}{2} \right],$$

where e_c and n_c are the number of edges and nodes respectively within cluster c and $\gamma > 0$ is a resolution parameter. This resolution parameter defines the internal density of the clusters such that: 1, no two clusters can be merged to lower $\mathcal{H}$; and 2, no cluster can be split to lower $\mathcal{H}$. Finally, the partitioning is refined to make all clusters well-connected – that is, such that from each node in a cluster you can reach every other node without leaving that cluster. This tries to guarantee that clusters have smooth distributions of important genes, as opposed to disconnected clusters which can have multiple pockets of more self-similar cells and still resist splitting as they are all more unlike the cells around them.

To summarize the cells within each cluster, all genes whose mean expression is higher within a given cluster than across all other cells were considered as possible "markers." The significance of these markers was evaluated by a Mann-Whitney U test with Bonferroni correction. The resulting p-values and log fold-change were used to rank these cluster "markers" and the highest-ranked genes were visually inspected. A large dominance of mitochondrial genes, i.e. those starting with "mt-" in the annotation file, or erythrocyte genes, e.g. hemoglobin, was used as criterion for cluster-level removal.

A second round of filtering was then done with these now individually-filtered samples, normalizing and clustering all of the cells together. The intersection of the gene set was taken across the samples, yielding a matrix where every gene was measured in every sample. Similar criteria as above were used to eliminate clusters from the combined partitioning. This process was repeated until no clusters were dominated by mitochondrial or erythrocyte genes, dropping genes that were not measured in any cell still remaining.

The filtered cells were normalized and clustered all together, and "markers" were calculated for each cluster as above. These markers were manually inspected for cell type enrichment, using the Human Protein Atlas as a reference for immune subtypes and the markers from a single-cell investigation of the small intestine for epithelium- and goblet-like cells.[35, 9]

2.6 Displaying Expression Data

For displaying the expression of a given gene over a large number of cells, one commonly used plot is the violin plot, which marks the minimum and maximum as the endpoints of a centerline that splits a histogram-like density cloud. However, the density cloud gives no measure of magnitude between different plots, so violin plots are often plotted with a random subsection of the data points to imply absolute density. Here, there is an additional problem of the high fraction of zero for many genes leading to a visually uniform density cloud for all nonzero values, thereby masking the nonzero distribution. To compensate for this, violin plots throughout this thesis are shown with the zero values excluded and the percent zero marked beneath the plot.

Additionally, it is helpful in this context to show the means of the data – both each sample mean individually and the mean of sample means within each model. The mean of sample means within each model is used, as each sample is treated as a repeated measure of the model, with each cell within that sample being a point measured from that sample's particular distribution. The following section will use the same measures as marked in Figure 2-2 as summary statistics.

2.7 Renormalization & Prioritization

The data were then re-normalized from count data by dividing each value by the sum of counts for that cell and multiplying by 10^6 to yield CPM (counts per million). Finally, the data were arcsinh-transformed to expression values as $E = \text{arcsinh}(CPM) =$

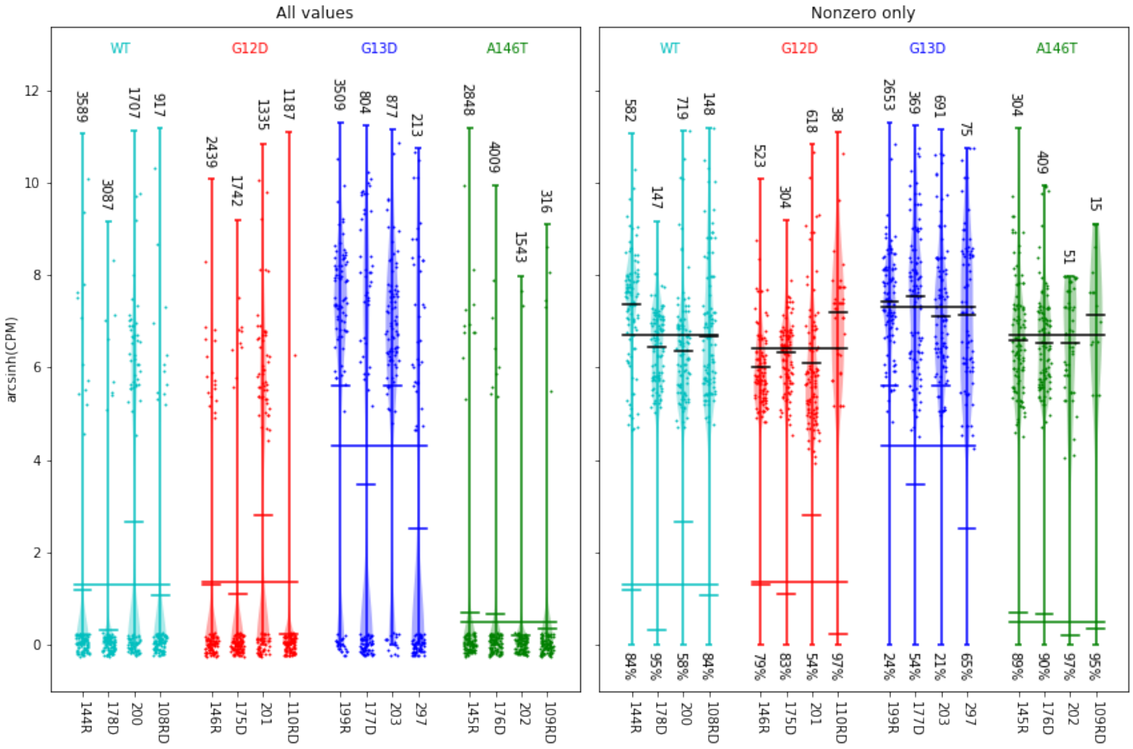

Figure 2-2: Description of violin plot format

Left: traditional violin plots

Right: modified to only show nonzero values

Marked on both plots are the sample mean, mean of samples means within each model, and number of nonzero cells. On the right, the percent nonzero and nonzero parallels to the means (in black) are also shown. Up to 100 random cells for each sample are plotted with small horizontal jitter; if there are fewer than 100 cells to display, all are shown. The expression of Mmp7 is used as an example.

$\ln\left(CPM + \sqrt{CPM^2 + 1}\right)$. This transformation was used instead of the GF-ICF normalization above since it does not depend on the distribution of expression across all cells, and therefore could be more suitable for studying distributions of values across a subset of cells.

To look for signals coming from the non-immune compartment that were significantly different in G13D compared to the rest of the models, the ligand-receptor interactions [29] were first limited to pairs where both the ligand and receptor were measured in every sample across the experiment. Then, these ligands' expressions across all cells within each model regardless of sample were used as repeated measurements in a Kruskal-Wallis H-test (the rank-based i.e. non-parametric extension of one-way ANOVA) with Benjamini-Hochberg correction to look for ligands that have significant differences across the four models. Next, these significantly changing ligands were then subset to those where the mean of the G13D sample nonzero means (i.e. the mean of only those cells in the sample were the gene was detected) were an extremum of the equivalent mean of means for the other models. Post-hoc tests were run by Mann-Whitney U with Benjamini-Hochberg correction on the sample means to determine how many of the other three models were different from the G13D samples. Depending on whether 1) any model, 2) at least two models, or 3) all three of the other models were significantly different from G13D, three lists of possibly interesting ligands were defined, each containing all of the ligands captured by the next smaller set. These are listed in Supplemental Table B.1.

For prioritization, three different methods were used: 1) consider only those ligands where the wingspan (i.e. $max - min$) of the model mean of means is larger than the wingspan of the sample means within each model individually; and 2) rank the absolute differences, or 3) percent differences between G13D as the extremum and its next nearest model mean of means. The first method prioritizes those genes with low variability of the sample means within a single model given large variability in the model mean of means. The second and third prioritize ligands by the separation of G13D as the extremum from its next nearest neighbor, whether by absolute or percent difference. These are all shown in Figure 2-3.

Three different subsets of expression were considered: 1) binarizing the data, replacing all nonzero values with one; 2) using only the nonzero data; and 3) using all the data directly. The mean of the first is equivalent to the fraction of cells expressing the gene, the second to the nonzero mean described previously, and the third to the traditional mean used in single-cell RNAseq.

For ligands of interest expressed on non-immune cells in G13D, the corresponding receptors were examined for significantly nonzero expression to find which cell type(s) could be receiving said signal. To do this, the fraction of nonzero for each cognate receptor was calculated for each cell type across the G13D model (regardless of sample),

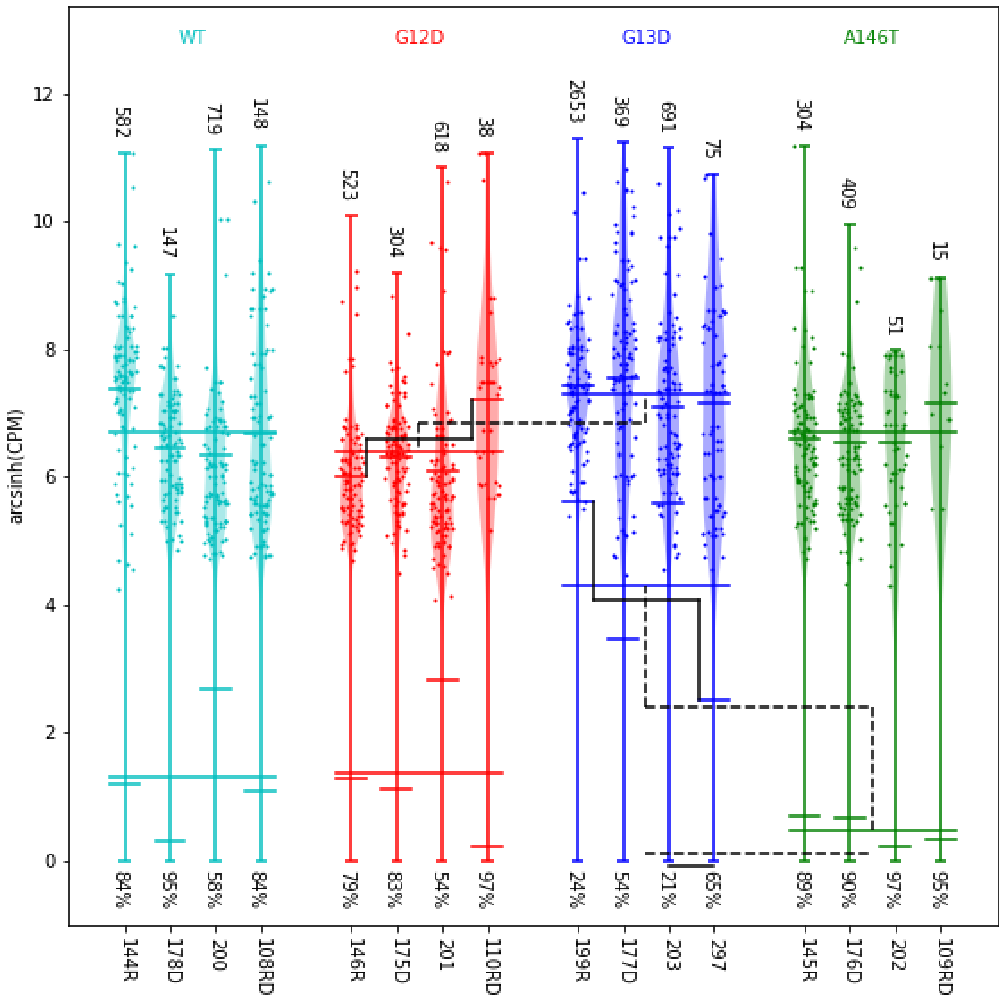

Figure 2-3: Visualization of wingspan tests
For each pair of solid (within model) and dashed lines (between models), the wingspans are:

Top: largest range of nonzero means

Middle: largest range of overall means

Bottom: largest range of percent zero

A comparison is significant if the dashed wingspan (between models) is larger than solid wingspan (within model). Here, Mmp7 is used as an example, which is significant by percent zero and overall mean.

then these values were binarized such that below some threshold fraction the receptor was considered “off” in that cell type and above such was considered “on” and capable of receiving a signal. This threshold was varied across the full range of possible values (0 to 1, inclusive) in increments of $0.01 = 1\%$. The receptors were then collapsed back onto their cognate ligand of interest and the number of cell types where at least one receptor was “on” for a given ligand was calculated. Repeating this for each threshold value yields a count of how many cell types could sense the ligand given a minimum required fraction of nonzero receptor expression that is monotonically decreasing – as the required fraction of nonzero receptor expression increases, the number of cell types that can pass that threshold decreases. The ligands of interest can then also be prioritized based on the highest nonzero fraction for any of its cognate receptors in at least one cell type, with a lower fraction zero implying a lower sensitivity for a homogeneous cell type.

Chapter 3

Results

3.1 Filtering, Visualization, & Broad Classification

After filtering, wild type (WT) samples produced $12,297$ cells, G12D $8,446$ cells, G13D $9,953$ cells, and A146T $10,206$ cells for a total of $40,902$ cells in this experiment (see Table 2.1). In turn, WT makes up 30% of the total cells, with G13D and A146T each contributing around 25%, and G12D filling the remaining nearly 20%.

After processing and filtering (see Sections 2.4 and 2.5), including two rounds of cluster-level filtering on the full dataset, the data were visualized using PCA-UMAP.[37, 24] Both model identity and Leiden clustering (with $\gamma = 1$) are shown in Figure 3-1.[34] Different regions show variable intermixing of the different models, e.g. cluster 12 is solely G12D while cluster 2 is well-mixed between all four models.

In the classification step (see Section 2.5), clusters 2, 4, 10, 14, and 16 were identified as immune cells, while the remaining clusters showed minimal immune / hematopoietic markers. These clusters were taken as the immune compartment, and the immune and non-immune compartments were separately clustered and visualized, and had markers calculated for further classification. The results of this round of clustering are shown in Figure 3-2. G13D contributed over 40% of the immune cells, with WT following it at nearly 30% – only 15% came from each of G12D and A146T. In contrast, WT and A146T each contributed about 30% of the non-immune cells, with G12D and G13D each adding 20%.

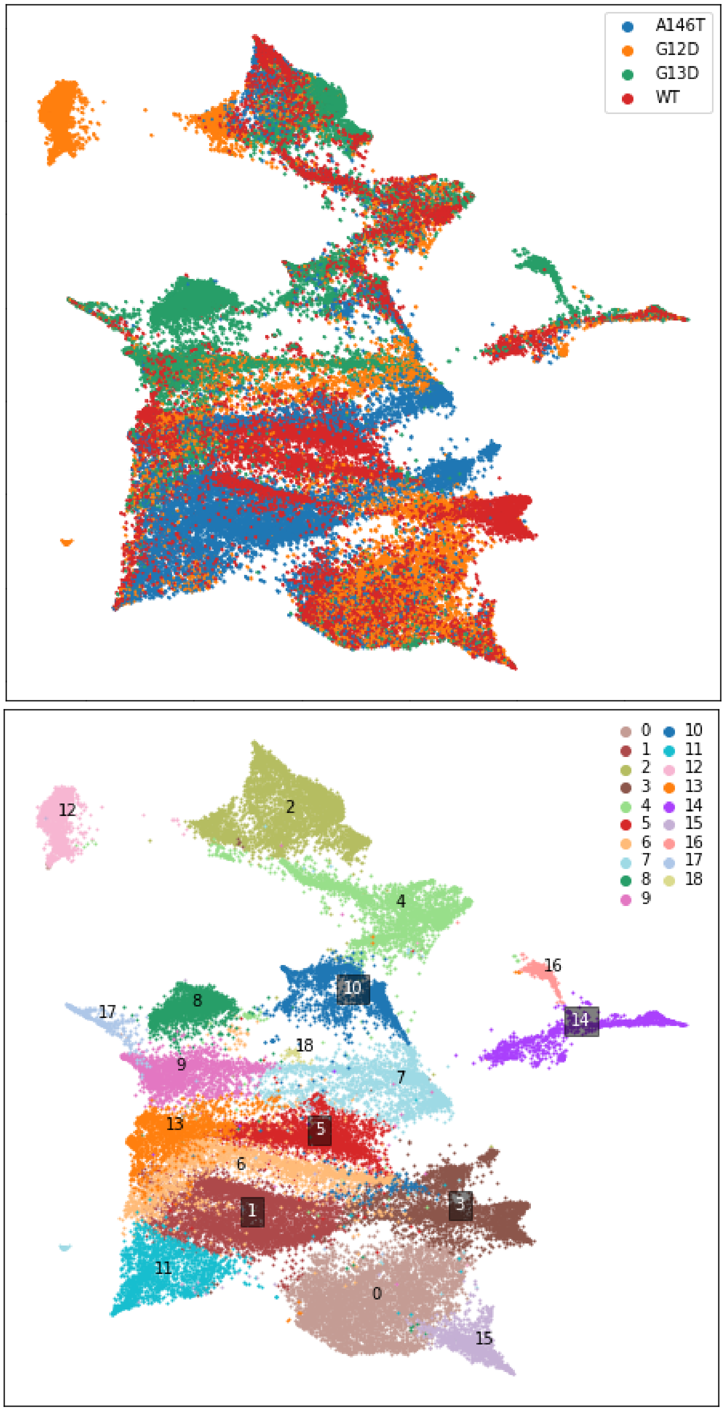

Figure 3-1: PCA-UMAP visualization of full dataset using the first 50 PCs and $\sqrt{n_{cells}}$ neighbors

Top: which model each cell came from

Bottom: Leiden clustering with $\gamma = 1$

When considering only the twelve samples (three per model) that were not enriched for any marker, G13D samples are much more dominated by immune cells than the other three models. In fact, two of the three unenriched G13D samples are more than half immune, while this is true for no other unenriched sample (see Table 3.1). Furthermore, there are only three other samples that are more heavily dominated by immune cells than the third unenriched G13D sample.

3.2 Further Classification

For the immune compartment, this second round of classification found well-defined cell type identifications for two-thirds of the resulting clusters. Clusters 7 and 12, comprising 9.4% of immune cells, were dominated by epithelial and mucosal genes, and so were classified as "Colonic" contaminants in the immune compartment. Clusters 5, 6, and 9, comprising 19.4% of immune cells, were of unclear cell type, and so were considered a single heterogenous "Unknown" cell type for downstream analysis. All five of these clusters lie either between the macrophages (clusters 0, 10) and granulocytes (clusters 1, 2, 11, 13) or on the outskirts as nearly isolated clusters alongside the plasma cells (cluster 8). Both B cells (cluster 3) and combined T cells / NK cells (cluster 4) segregated from the remaining cells. Finally, two types of dendritic cells were seen – conventional (cluster 14) and plasmacytoid (cluster 15).

Within the mixed T and NK cell cluster, a further round of normalization, clustering, and classification was attempted, but clear internal divisions could not be drawn and so were treated as a single cell type (see Figure B-1).

In the non-immune compartment, however, there was no clear subtyping based on markers despite the mix of cell types expected to be present – including endothelial, secretory, healthy epithelial, and tumor. Clusters instead seemed to be majorly defined by the fraction of each model they derive from, with 3 of the 15 clusters comprising 16.4% of the cells containing essentially a single model and a further 8 clusters containing a single model that accounts for the majority. In many cases, these clusters derive majorly from only a small subset of the samples. As there was no obvious

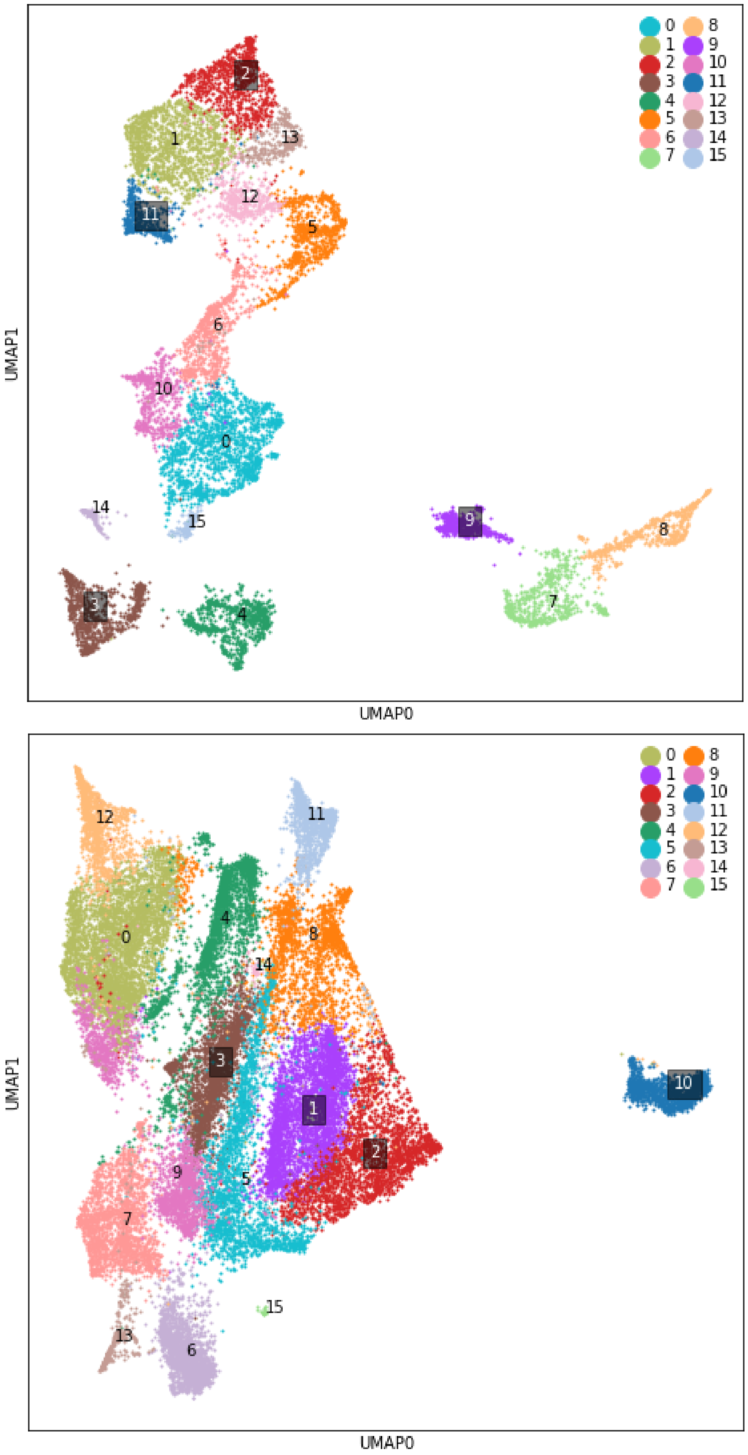

Figure 3-2: PCA-UMAP visualization of immune and non-immune compartments using the first 50 PCs and $\sqrt{n_{cells}}$ neighbors, with $\gamma = 1$

Top: Leiden clustering of the immune compartment

Bottom: Leiden clustering of the non-immune compartment

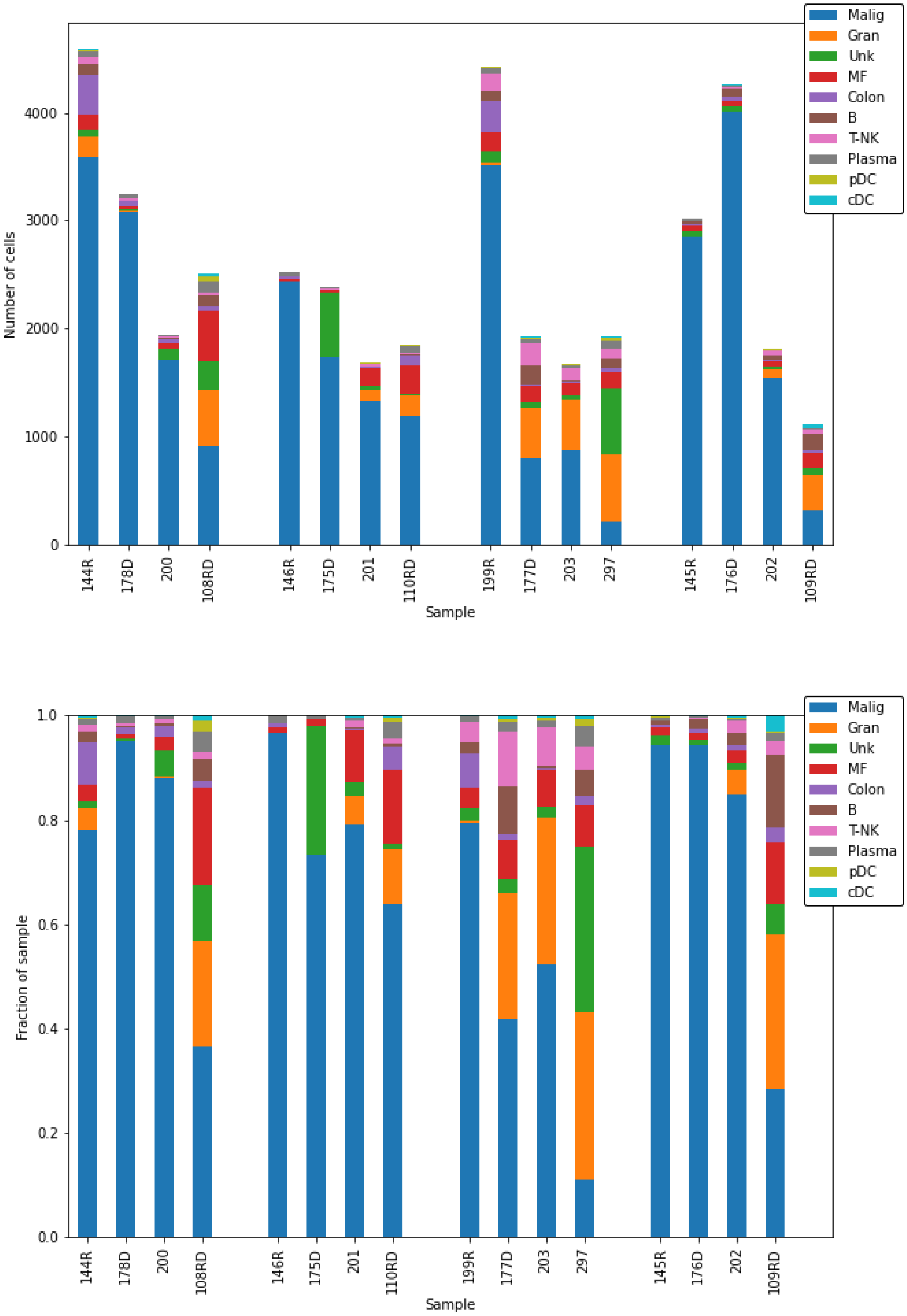

Figure 3-3: Number and fraction of cell types by sample

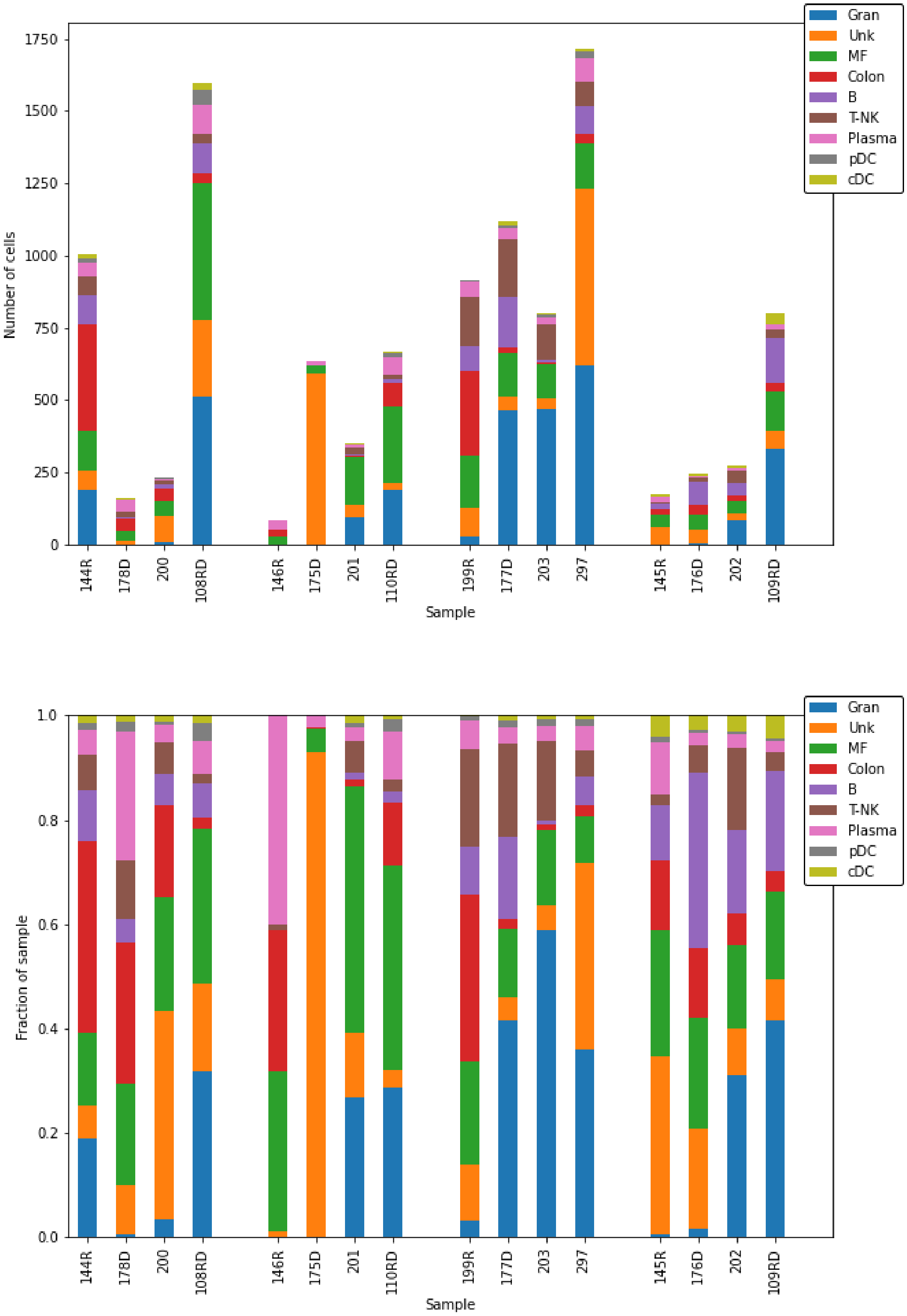

Figure 3-4: Number and fraction of cell types by sample in the immune compartment

way to separate these cells further for downstream analysis, all non-immune cells were grouped as a single “non-immune” cell type, disregarding internal variation and likely lowering power compared to using truly cell type-pure groupings.

The final classification percentages are shown in Figures 3-3 and 3-4 and Tables 3.1 and 3.2. As expected, the CD45-enriched samples (in *italics*) have a higher immune content than their unenriched counterparts – and in fact the CD45-enriched sample with the lowest immune fraction is still only surpassed by a single unenriched outlier (sample 203). Secondarily, it can also be noted that the G13D samples have higher proportion of immune cells than those with different KRas alleles.

3.3 Ligands of Interest in G13D

To try to uncover why the G13D samples have higher immune burden, we can focus on the ligands expressed by the non-immune cells that could be sending signals out into the tumor's microenvironment. For use with extant methods, such as iTALK and CellPhoneDB, several assumptions were evaluated on the arcsinh-normalized data.

iTALK considers either the top half of genes by average expression, or differentially expressed genes (DEGs) as determined by a number of possible methods.[36] One option that is implemented for identifying DEGs is a simple univariate Wilcoxon ranked-sum test, equivalent to the Mann-Whitney U used as a post-hoc test here. However, we wanted a method that takes advantage of the repeated measures we had for each model (i.e. three mice per genotype) and that is able to compare between more than two groups.

CellPhoneDB only considers a gene if it was detected (i.e. nonzero) in more than 10% of its cluster.[5] As shown in Figure 3-5 however, this would restrict analysis to very few ligands and receptors – and an even smaller fraction of interactions given that CellPhoneDB requires both the ligand and the receptor to pass this threshold to calculate a score. Due to the preponderance of zeros in this dataset, this threshold was considered too restrictive, and led to consideration of the relative contributions of nonzero values and the fraction nonzero to the mean over all of the cells.

Table 3.1: Fractions of each sample assigned to each cell type
Italics: CD45-enriched samples; directly followed by their corresponding unenriched sample – means no cells were found in this particular sample / cell type combination.

Percentage of Sample	**WT**				**G12D**				**G13D**				**A146T**			
	108RD	144R	178D	200	*110RD*	146R	175D	201	*297*	199R	177D	203	*109RD*	145R	176D	202
Non-immune	*36.4*	78.1	95.1	88.0	*64.0*	96.6	73.2	79.1	*52.3*	79.3	41.9	11.0	*28.3*	94.3	94.2	85.0
Immune	*63.6*	21.9	4.9	12.0	*36.0*	3.4	26.8	20.9	*47.7*	20.7	58.1	89.0	*71.7*	5.7	5.8	15.0
Granylocyte	*20.3*	4.1	0.0	0.4	*10.3*	–	–	5.6	*32.1*	0.7	24.2	28.0	*29.7*	0.0	0.1	4.7
Unknown	*10.7*	1.4	0.5	4.8	*1.2*	0.0	24.9	2.6	*31.7*	2.2	2.6	2.3	*5.7*	2.0	1.1	1.3
Macrophage	*18.8*	3.0	1.0	2.6	*14.2*	1.0	1.2	9.8	*8.1*	4.1	7.7	6.9	*12.0*	1.4	1.2	2.4
Colonic	*1.3*	8.1	1.3	2.1	*4.4*	0.9	0.0	0.3	*1.8*	6.6	1.1	0.5	*2.9*	0.8	0.8	0.9
B cell	*4.1*	2.1	0.2	0.7	*0.7*	–	–	0.3	*4.9*	1.9	9.1	0.4	*13.7*	0.6	1.9	2.4
T/NK cells	*1.3*	1.5	0.6	0.7	*0.9*	0.0	0.0	1.2	*4.5*	3.9	10.5	7.3	*2.7*	0.1	0.3	2.4
Plasma cells	*4.0*	1.0	1.2	0.4	*3.3*	1.3	0.6	0.5	*4.1*	1.1	1.8	1.4	*1.5*	0.6	0.1	0.4
pDC	*2.0*	0.3	0.1	0.1	*0.8*	–	–	0.2	*1.3*	0.1	0.7	0.6	*0.4*	0.1	0.0	0.1
cDC	*1.0*	0.3	0.1	0.2	*0.3*	–	–	0.3	*0.6*	0.0	0.6	0.3	*3.0*	0.2	0.2	0.4

Table 3.2: Fractions of each sample assigned to each cell type in the immune compartment
Italics: CD45-enriched samples; directly followed by their corresponding unenriched sample
– means no cells were found in this particular sample / cell type combination.

Percentage of Immune	**WT**				**G12D**				**G13D**				**A146T**			
	108RD	144R	178D	200	*110RD*	146R	175D	201	*297*	199R	177D	203	*109RD*	145R	176D	202
Granulocyte	*31.9*	18.9	0.6	3.4	*28.6*	–	–	26.7	*36.1*	3.2	41.6	58.8	*41.4*	0.6	1.6	31.1
Unknown	*16.8*	6.4	9.4	39.9	*3.4*	1.2	93.1	12.5	*35.6*	10.7	4.4	4.8	*8.0*	34.1	19.2	8.8
Macrophage	*29.6*	13.9	19.5	21.9	*39.3*	30.6	4.4	47.2	*9.1*	19.8	13.2	14.5	*16.8*	24.3	21.2	16.1
Colonic	*2.1*	36.9	27.0	17.6	*12.1*	27.1	0.2	1.4	*2.0*	32.0	1.9	1.0	*4.0*	13.3	13.5	5.9
B cell	*6.5*	9.5	4.4	6.0	*1.9*	–	–	1.4	*5.5*	9.2	15.7	0.9	*19.1*	10.4	33.5	16.1
T/NK cells	*2.1*	6.9	11.3	6.0	*2.4*	1.2	0.2	6.0	*5.0*	18.9	18.0	15.4	*3.8*	2.3	5.3	15.8
Plasma cells	*6.3*	4.8	24.5	3.4	*9.3*	40.0	2.2	2.6	*4.6*	5.5	3.1	2.9	*2.1*	9.8	2.4	2.6
cDC	*3.2*	1.3	1.9	0.4	*2.2*	–	–	0.9	*1.5*	0.7	1.2	1.3	*0.5*	1.2	0.4	0.7
pDC	*1.6*	1.5	1.3	1.3	*0.7*	–	–	1.4	*0.6*	0.2	1.0	0.6	*4.3*	4.0	2.9	2.9

This was further noted when looking at the distributions of individual ligands and receptors. Figure 3-6 shows the distribution and relative magnitude of a sampling of the 50 highest-expressed ligands and receptors by overall mean of the 601 that were measured along with at least one counterpart. All of these have a peak at zero alongside a normal-looking distribution for the nonzero values. This led to consideration of nonzero values and the fraction of these values as separate measures that both contribute to the overall mean. Mathematically,

$$\begin{aligned} mean_{overall}\left(e\right) &= \frac{\sum_i e_i}{n}, \\ mean_{nonzero}\left(e\right) &= \frac{\sum_i e_i}{\sum_i \mathbb{1}\left(e_i \neq 0\right)}, \text{ and} \\ frac_{nonzero}\left(e\right) &= \frac{\sum_i \mathbb{1}\left(e_i \neq 0\right)}{n}, \text{ therefore} \\ mean_{overall}\left(e\right) &= mean_{nonzero}\left(e\right) * frac_{nonzero}\left(e\right) \end{aligned}$$

where $\mathbb{1}\left(x\right) = 1$ if x is true and 0 otherwise. Note that it is possible that the product of two components each with no significant differences between/across models can itself be significantly different.

Looking for ligands on non-immune cells that are different in G13D compared to the other models, lists were generated that prioritized these ligands using both of these components as well as the overall mean (see Section 2.7; Supplemental Table B.1). Using a wingspan test – where the largest range of mean expression within any single model has to be smaller than the range of the model mean expressions – using these metrics, 10 ligands were found to be of top priority: B2m, Plat, Vim, Fn1, Apoe, Mmp7, Agrn, Pgf, Sema3e, and Efna5. These ligands were also found near the top of the prioritization lists when ranking by difference between G13D and the next most extreme model, using the same array of statistics. Two further ligands were found in two of the wingspan tests using the nonzero mean where the mean of G13D sample means is the minimum of any model, and so were included in the list of top priority ligands even though they were not found to be high priority using other metrics.

Furthermore, the prioritization lists by difference also relatively consistently found

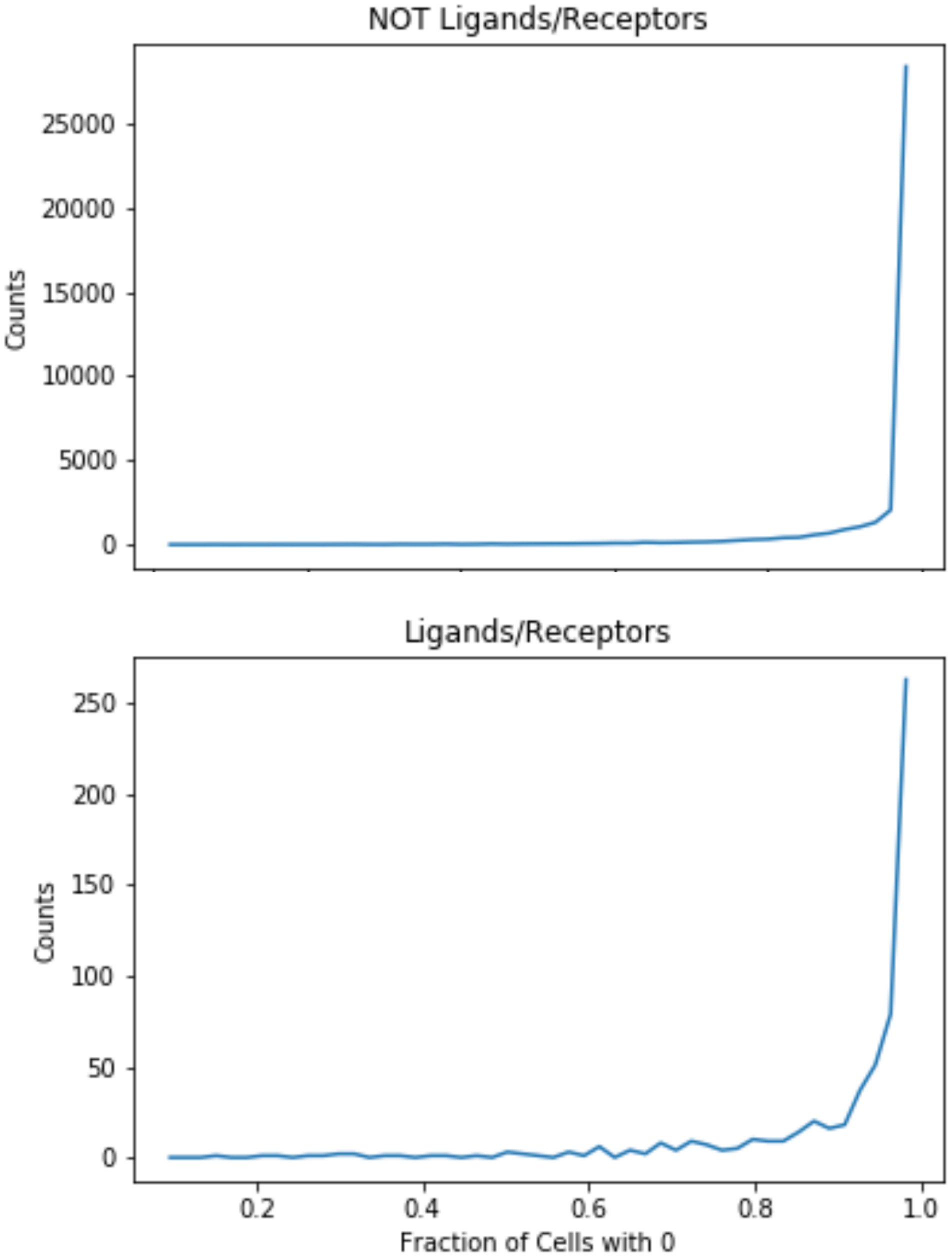

Figure 3-5: Histograms of fraction nonzero for genes and ligands/receptors

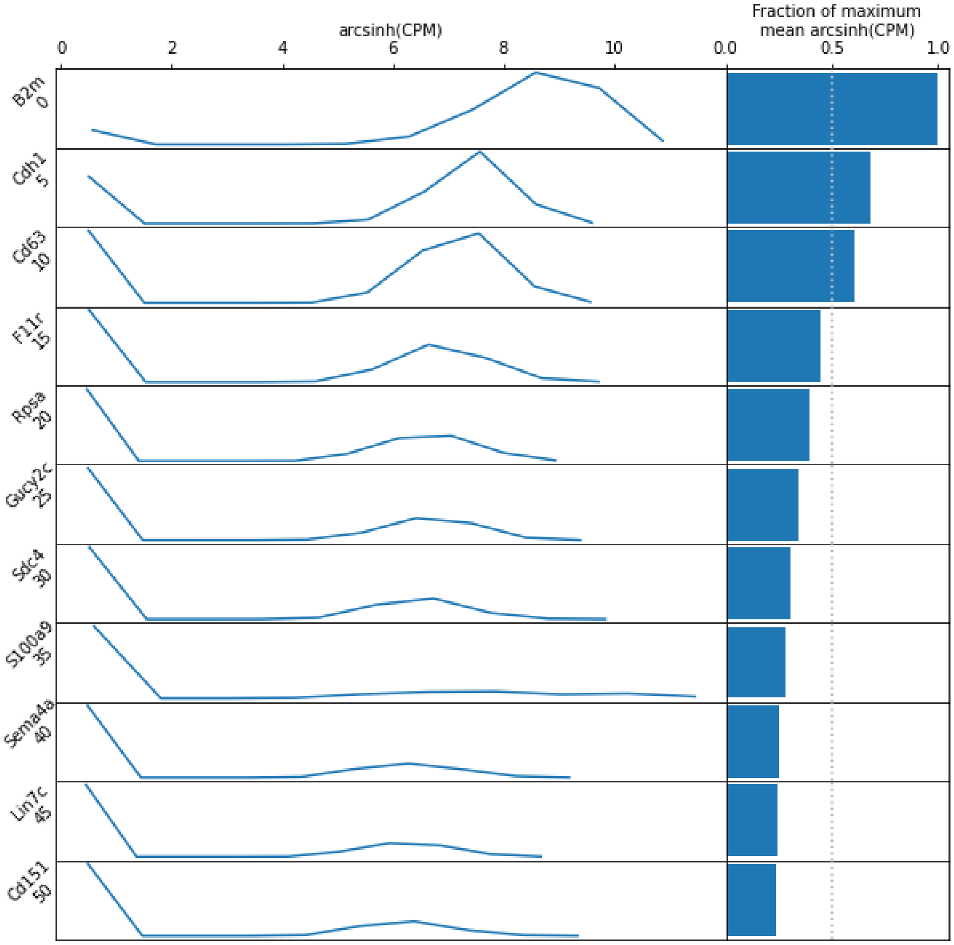

Figure 3-6: Distributions of highest-expressed ligands and receptors

Left: the distribution of arcsinh-normalized expression

Right: the fractional expression relative to the maximum mean

Every fifth ligand/receptor is shown when ranked by overall mean for the first 51 highest-expressed ligands/receptors. The number under the gene name is its 0-indexed rank by overall mean.

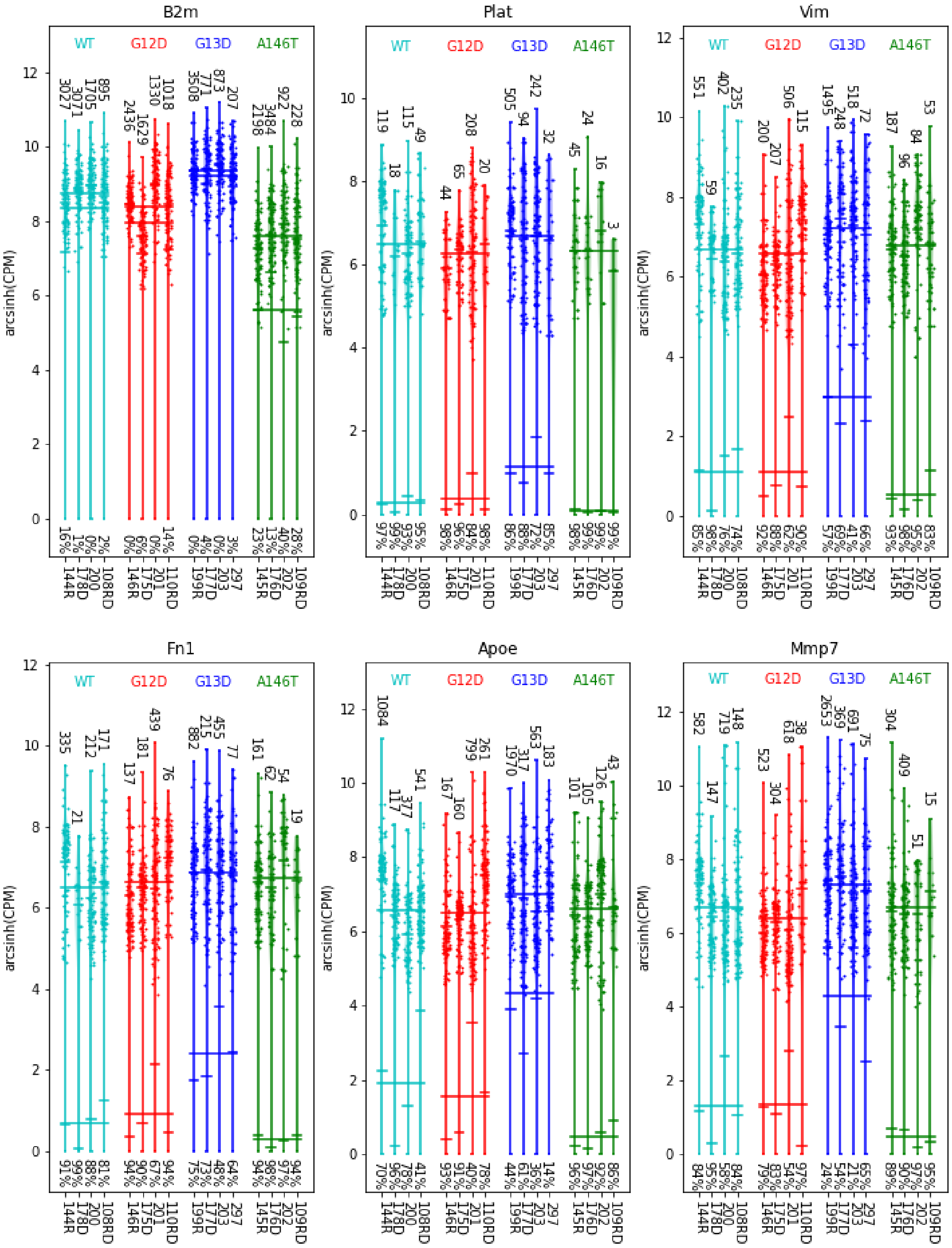

Figure 3-7: Expression of ligands of interest in non-immune cells

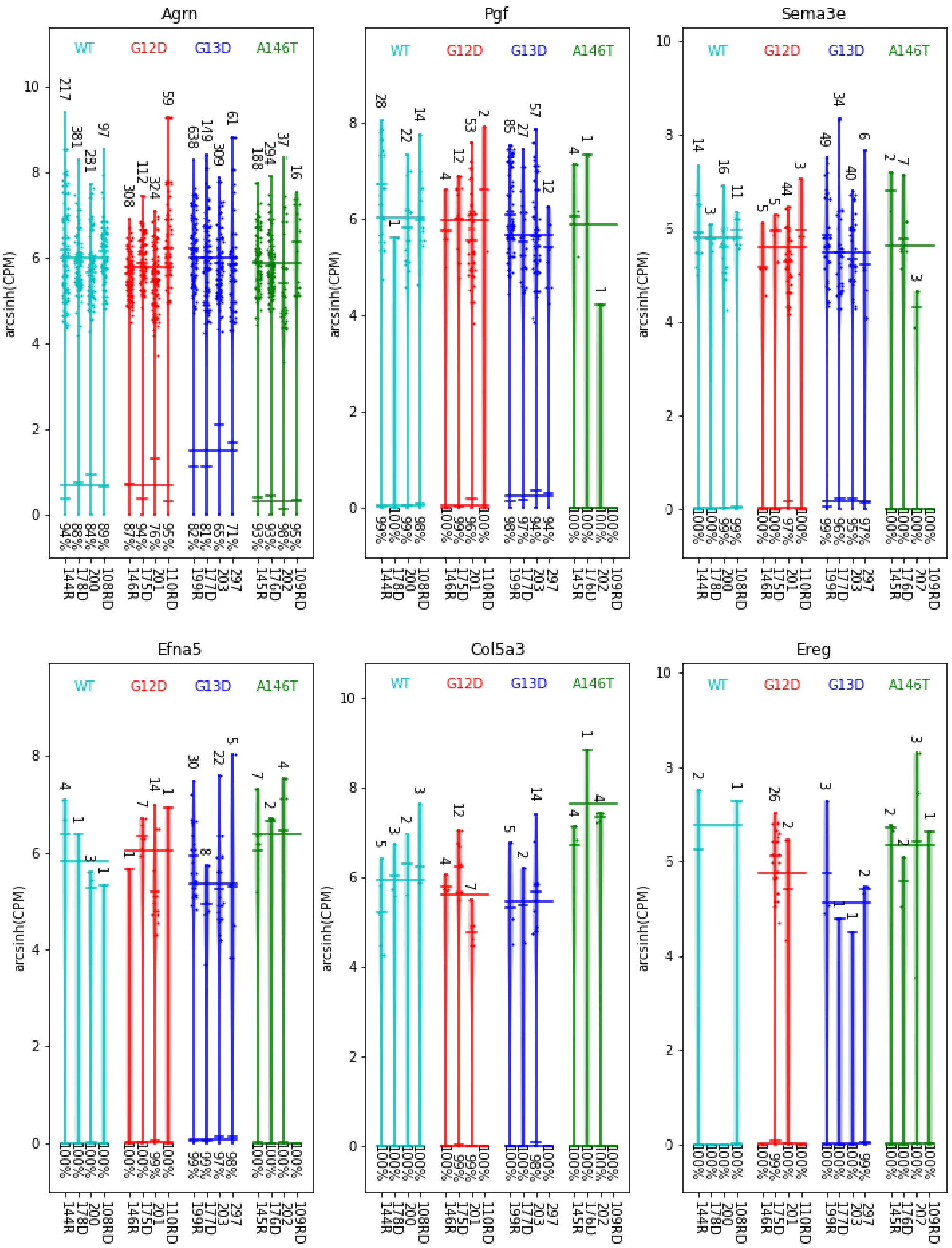

Figure 3-7: Expression of ligands of interest in non-immune cells, cont'd

Table 3.3: Prioritization of ligands of interest using their cognate receptors

* denotes ligands that were deprioritized, as their expression in the immune compartment likely dominates over that from non-immune cells (see Figure 3-9) The threshold value is the fraction of nonzero receptor expression required for receiving a ligand's signal for which no cell type has a single receptor that passes and therefore is the threshold above which the ligand cannot affect any cell type.

Gene	**Threshold**
Vim	0.85
Mmp7	0.85
Fn1*	0.85
Apoe*	0.57
Plat	0.46
Pgf	0.29
Efna5	0.26
Sema3e	0.25
B2m	0.21
Agrn	0.19
Col5a3	0.09
Ereg*	0.07

a handful of other genes – e.g. Ptn and Sema3c – but these were not included for further prioritization as it was unclear how to handle the inconsistencies between different metrics while the wingspan test produced a reasonable number of candidate ligands to follow-up on experimentally. Delving into these lists is left to those readers who want to dig deeper into the nuanced biology of these genes in the current context.

For these 12 ligands of interest, the expression levels of their cognate receptors across G13D samples were checked to see which cell types (if any) could be receiving these differential signals. The number of possible receiver cell types was plotted for a range of threshold nonzero receptor fractions, treated as a boundary for where the cell type as a homogeneous whole would be able to receive the signal (see Section 2.7). It is good to keep in mind that this is not a wise interpretation for cell types that have a lot of internal variation – and is especially bad for the Colonic and Unknown pseudo-cell types that were identified in the immune compartment which have multiple separable components. These curves are shown in Figure 3-8 and the resulting ordering in Table 3.3.

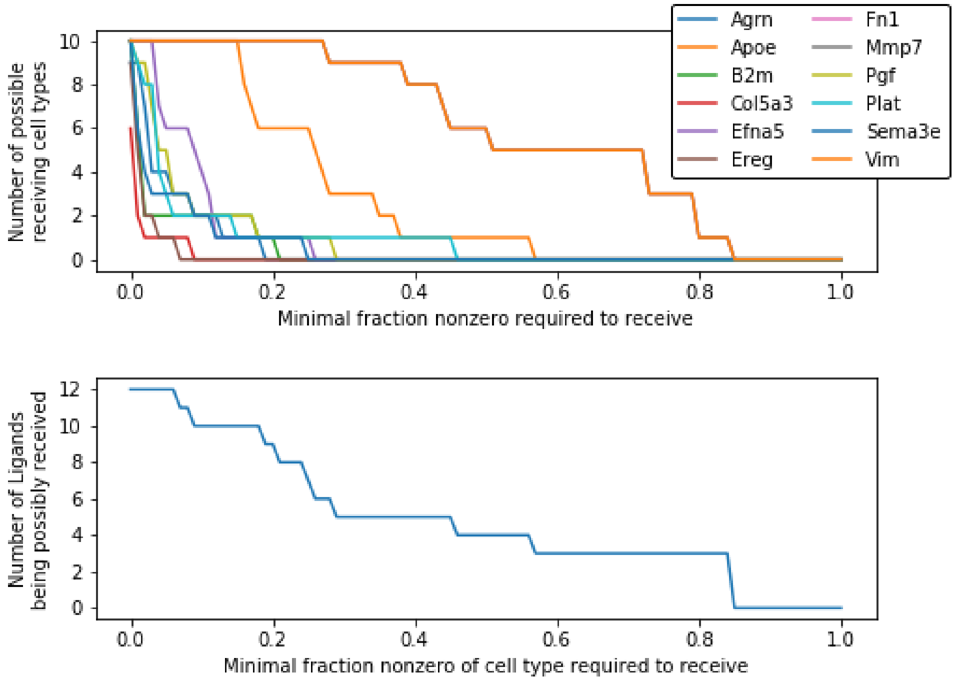

Figure 3-8: Prioritizing ligands of interest using their cognate receptors

Top: the number of cell types with at least one receptor for the ligand
whose fraction nonzero is greater than the threshold
Note: Mmp7 and Fn1 match Vim for their entire length
Bottom: the number of ligands that have at least one cell type that pass
the fraction nonzero threshold for at least one of their receptors

The x-axis is the threshold nonzero fraction above which the receptor is “on” across a given cell type and capable of receiving the signal of ligand presence.

The GF-ICF [8] expression of each ligand of interest was painted onto a PCA-UMAP projection [37, 24], coloring each cell with the mean of its cluster [34] to increase the prominence of the coloration (see Figure 3-9). Fn1, Apoe, and Ereg are mostly restricted to the immune compartment but are captured in the above analyses of non-immune cells. Their expression by non-immune cells will likely be overwhelmed by their immune expression and, as such, were deprioritized. Vim and Pgf are mostly expressed in the immune compartment as well, but a single non-immune cluster also has notable expression. This cluster also dominates the expression of Plat, Agrn, Sema3e, and Efna5. Additionally, if thinking of Mmp7 expression as a marker of Paneth cells, the Paneth-like cluster also express Col5a3. Lastly, B2m shows broad expression across all immune and some non-immune clusters.

Using the set of ligand-receptor interactions in [29], the set of receptors that interact with any of the ligands of interest was collated (see Table 3.4) and their expressions were also painted onto the PCA-UMAP projection, with each cell being colored by the mean of its cluster as above (see Figure 3-10). All receptors are shown regardless of expression level, as a threshold for determining significant cell type-wide response is not established. Five of the 12 ligands have only a single receptor measured in every sample, while Fn1 is by far the most prolific with 19 different receptors (dominated by the integrin family).

Some receptors – Atp1a3, Hfe, Itga2b, Itga4, Itga5, Itga8, Itga9, Itgb3, Lrp1, Lrp8, Nrp1, Plaur, Plxnd1, Sdc3, and Sorl1 – were mainly restricted to the immune compartment, while others – Cd151, Egfr, Epha4, Ephb2, Ephb6, Erbb4, Flt1, Itga2, Itga3, Itgb6, Ldlr, Lrp4, Lrp5, Nt5e, Sdc2, Tnfrsf11b, and Vldlr – were highly dominated by the non-immune compartment. The remaining receptors were spread across both subsets of cells. This segregation of expression leads to the interesting possibility that a single ligand could be causing different signals based on which receptor is expressed in different cell types.

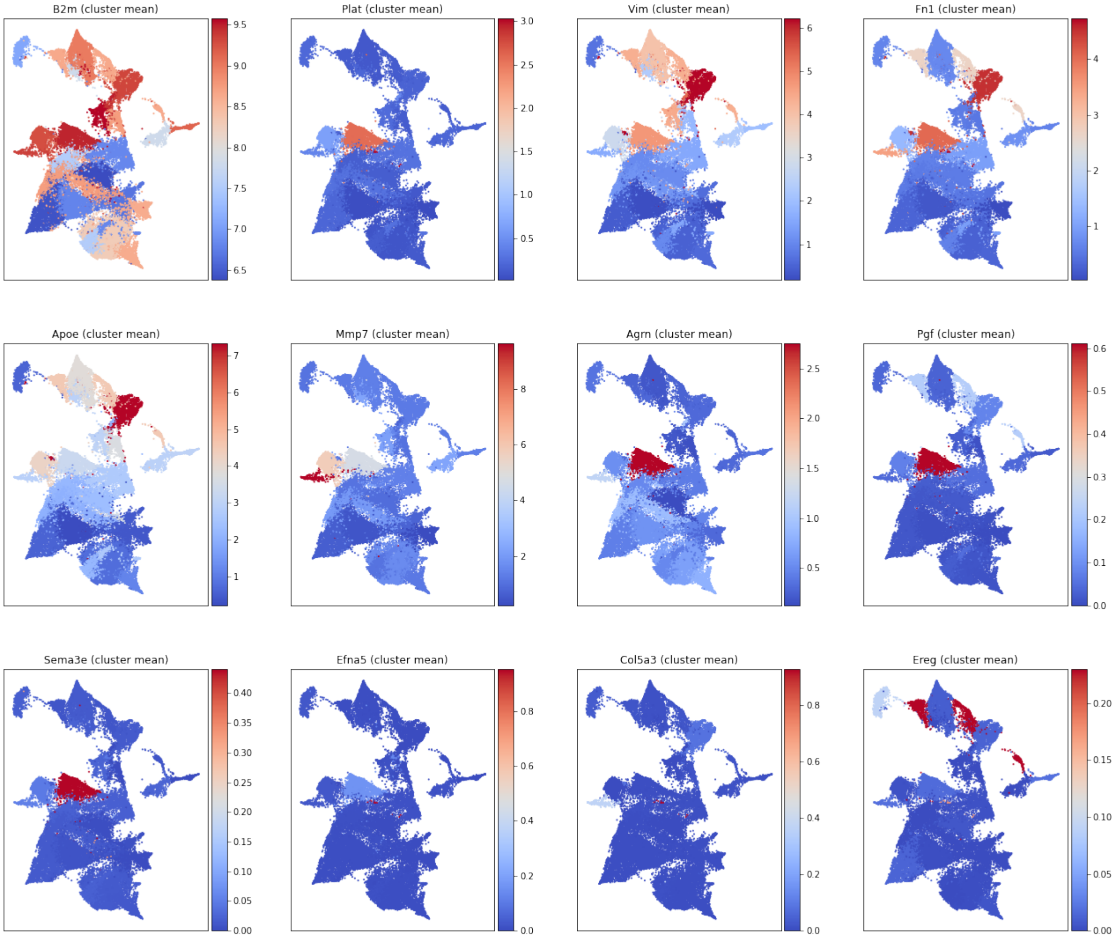

Figure 3-9: Ligand of interest expression in full dataset using the first 50 PCs and $\sqrt{n_{cells}}$ neighbors
Each cell is colored by the GF-ICF mean of its cluster (with $\gamma = 1$).

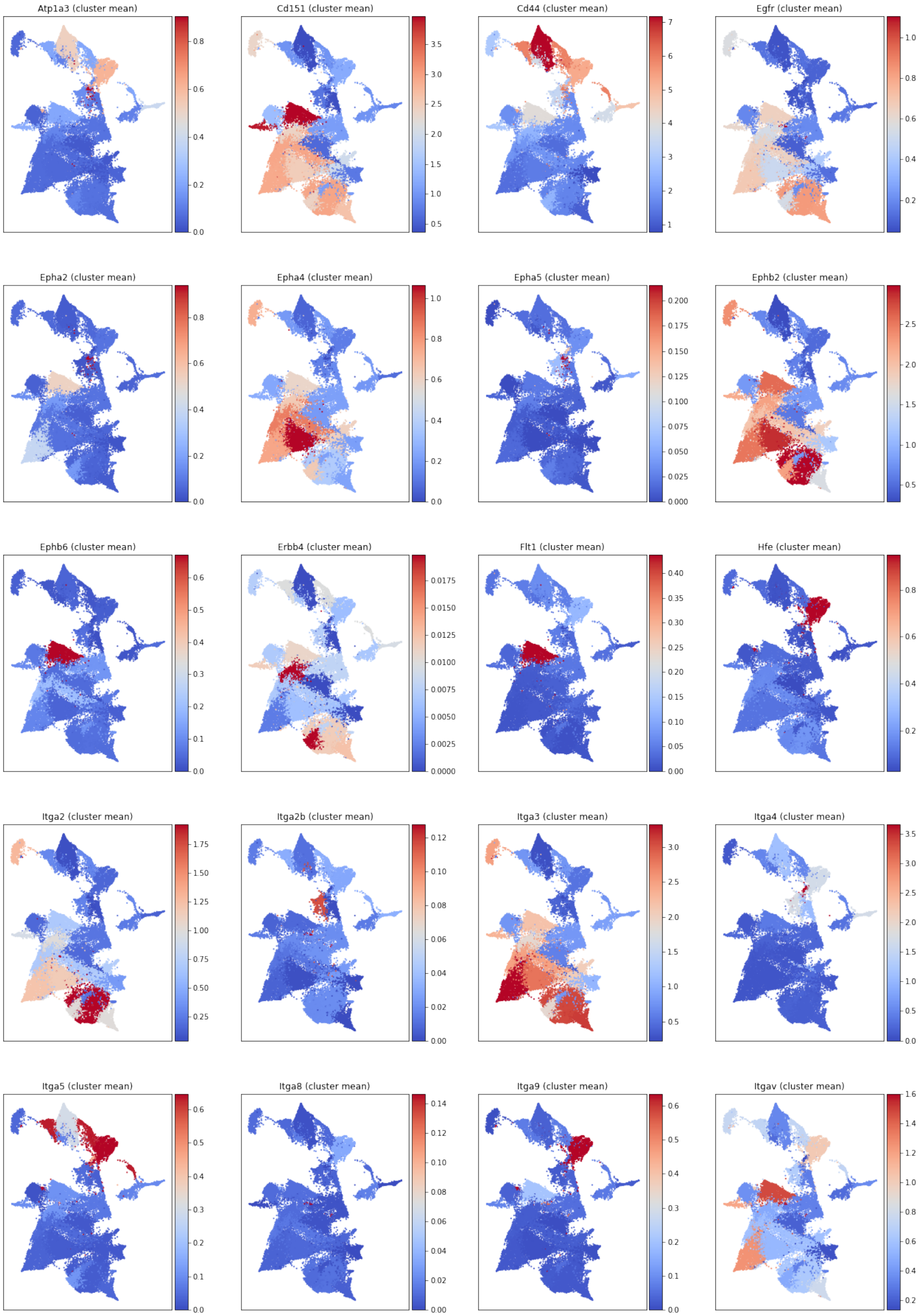

Figure 3-10: Receptors for ligand of interest expression in full dataset using the first 50 PCs and $\sqrt{n_{cells}}$ neighbors
Each cell is colored by the GF-ICF mean of its cluster (with $\gamma = 1$).

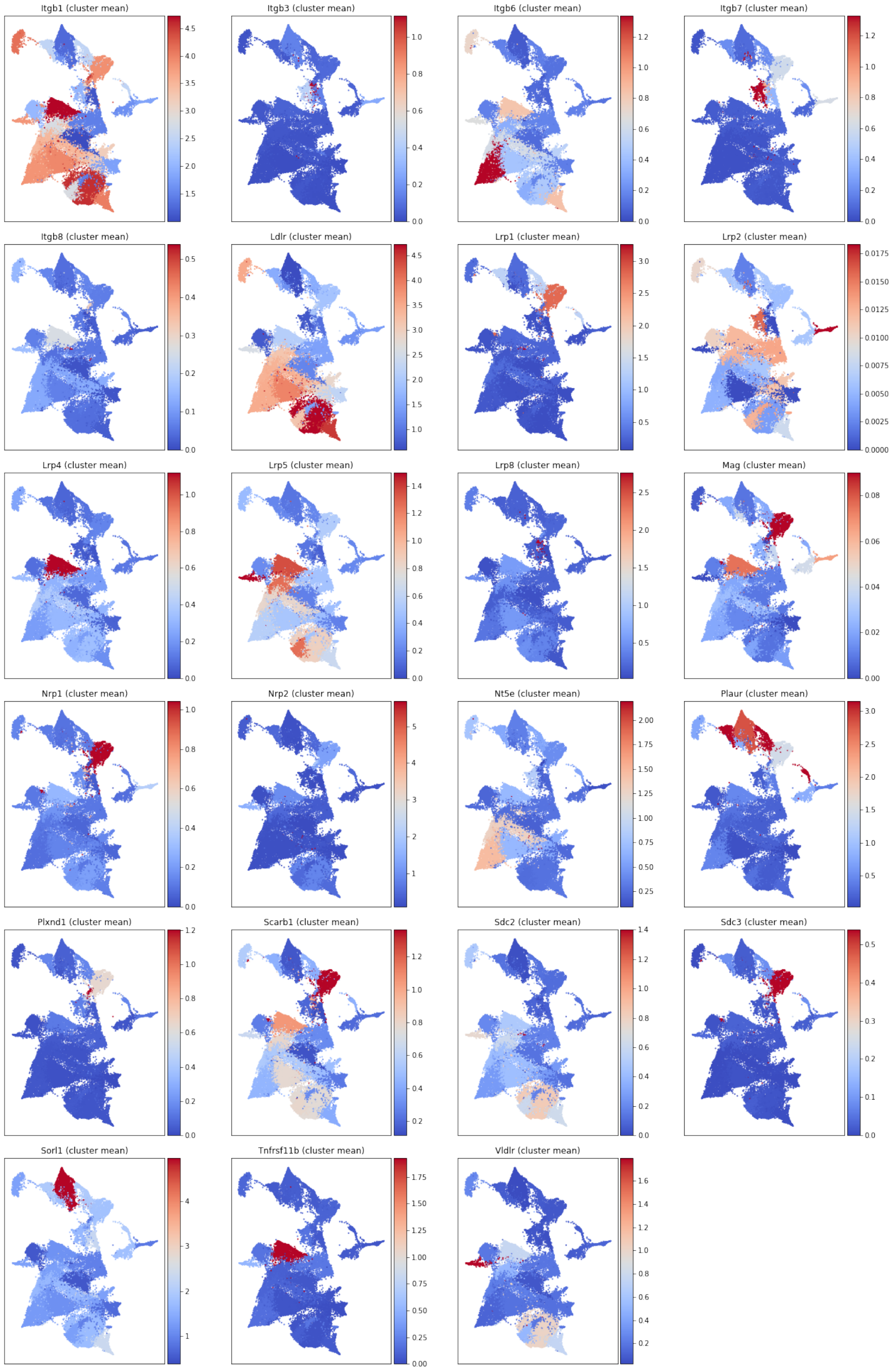

Figure 3-10: Receptors for ligand of interest expression in full dataset, cont'd

Table 3.4: Cognate receptors for ligands of interest

Ligand	**Receptor(s)**
B2m	Hfe
Col5a3	Sdc3
Plat	Lrp1
Sema3e	Plxnd1
Vim	Cd44
Agrn	Atp1a3, Lrp4
Ereg	Egfr, Erbb4
Mmp7	Cd151, Cd44, Erbb4
Pgf	Flt1, Nrp1, Nrp2
Efna5	Epha2, Epha4, Epha5, Ephb2, Ephb6
Apoe	Ldlr, Lrp1, Lrp2, Lrp5, Lrp8, Scarb1, Sorl1, Vldlr
Fn1	Cd44, Itga2, Itga2b, Itga3, Itga4, Itga5, Itga8, Itga9, Itgav, Itgb1, Itgb3, Itgb6, Itgb7, Itgb8, Mag, Nt5e, Plaur, Sdc2, Tnfrsf11b

Chapter 4

Discussion

4.1 Ligands of interest

Previous studies on differences between different KRas mutations have found only a few of the same ligands as were of interest here. Vim and Fn1 were found to be highly expressed by G13D-transformed human breast epithelium as compared to the same cell line transformed instead with G12D.[33] Similarly, the expression of these two genes was decreased by siRNA knockdown of KRas in both G13D-mutant MDA-MB-231 cells as well as KRas-wild type BT-549 cells, both human breast cancer lines.[16] What is known about others can even be somewhat contradictory. Low B2m expression, for example, was shown to be associated with recurrence of CRC and so results in a poor prognosis,[2] implying that G13D-mutants – with their high levels of B2m seen here – are less likely to recur than the other KRas alleles. Conversely, several studies have found that patients with G13D-mutant KRas have a worse prognosis than those with other KRas mutations.[31, 6]

Most of the ligands found here, however, are relatively novel for their importance in different KRas mutations. While Mmp7 has been shown to be upregulated by G12D-mutation of KRas in pancreatic ductal adenocarcinoma,[7] it appears that no one has published a comparison of Mmp7 expression across a series of KRas-mutant tumors. Ereg has been shown to be downregulated by KRas mutation compared to KRas-wild type tumors, and furthermore loses its positive correlation with clinical outcomes with mutation of KRas,[14] but that paper does not state which mutations were seen in their KRas-mutant group.

Overall, the importance of the ligands identified here is uncertain as there is not enough existing literature to support or refute most of them.

4.2 Strengths & Weaknesses

To begin filling the hole in the literature, this work attempts to be a first pass at studying the effects of KRas mutations on intercellular communication. Using four models with an n of only 3 mice (or 4 if counting the CD45-enriched sample separately), there is clearly much room to expand on this type of analysis. Additionally, several different prioritization methods were proposed here, but an evaluation of which set produces the most relevant results in vivo still awaits experimental follow-up or orthogonal computational validation.

However, this method is quite extensible, allowing for optimization e.g. of the summary statistic in ranked prioritization, while being quite robust in other senses, capturing a consistent set of ligands using either the mean across all cells or dividing it into two analyses of nonzero mean and fraction nonzero. Additionally, this form of analysis can be used for scRNAseq data, bulk sequencing of sorted cells, or quantified RNA-FISH for specific ligands as it is theoretically agnostic to the source of input data. Even mass cytometry data could be treated in an analogous way, using protein-level measurements to more directly quantify surface-bound ligands instead of inferring protein level from mRNA expression. As a set of conceptually simple ideas (requiring statistical change with either larger inter- versus intra-model variation or large separation between the extremum and its nearest neighbor), minimal assumptions need to be made to pull out differences that can be used as a starting place for furthering correlative or mechanistic explanations.

Chapter 5

Conclusions

A lot is going on with KRas mutations that is not fully understood due to systemic deficiencies in the literature as a whole. While on the largest scale, different mutant KRas alleles are recognized to have distinct clinical implications, they are usually compared as a block with wild type KRas for molecular and cellular comparisons. Across four different models of APC-driven colorectal cancer with four common KRas alleles, scRNAseq is able to find statistically large differences in cell-cell communication that set the G13D mutation apart from G12D and A146T mutations and wild type KRas, despite the large range in clinical outcomes between KRas-wild type and KRas-mutant tumors. While the analytical methods described here require very little in the way of complex mathematics, their simplicity of interpretation allows for the possibility of deep understanding. It captures two ligands (Vim and Fn1) known to have differential effects with different KRas mutations, as well as a handful of novel genes whose roles in this context are not well established. With further investigation into a small set of high priority ligands, these differences can hopefully be corroborated with other experimental or computational evidence and be understood with biological insight within the highly variable context of colorectal cancer.

Appendix A

GF-ICF L^2-Normalization

This section will show that using L^2-normalization of GF-ICF values for each cell removes its dependency on the total counts of that cell.

Proof. From Section 2.5, to convert a count matrix f with N cells to its corresponding expression matrix e, for gene i and cell j:

$$\begin{aligned} gf_{ij} &= \frac{f_{ij}}{\sum_i f_{ij}} \\ icf_i &= \begin{cases} log\left(\frac{N}{n_i}+1\right) & n_i \neq 0 \\ 0 & n_i = 0 \end{cases} \\ e_{ij} &= gf_{ij} * icf_i \,, \end{aligned}$$

where $n_i = |\{j : f_{ij} > 0\}|$ is the number of cells in which at least one UMI from the given gene was seen. Note that $0 \leq n_i \leq N$. For reference, the total counts of cell j is $\sum_i f_{ij}$. From substituting the first equation above in the last,

$$e_{ij} = \frac{f_{ij} * icf_i}{\sum_i f_{ij}}.$$

With L^2 normalization by cell, you instead have

$$\begin{aligned} e'_{ij} &= \frac{e_{ij}}{\sqrt{\sum_i e_{ij}^2}} \\ &= \frac{f_{ij} * icf_i / \sum_i f_{ij}}{\sqrt{\sum_i \left(f_{ij} * icf_i / \sum_i f_{ij}\right)^2}}. \end{aligned}$$

As $\sum_i f_{ij}$ does not depend on i, it can be pulled out of the summation.

$$\begin{aligned} e'_{ij} &= \frac{f_{ij} * icf_i / \sum_i f_{ij}}{\sqrt{\left(\sum_i f_{ij}\right)^{-2} \sum_i \left(f_{ij} * icf_i\right)^2}} \\ &= \frac{f_{ij} * icf_i / \sum_i f_{ij}}{\left(\sum_i f_{ij}\right)^{-1} \sqrt{\sum_i \left(f_{ij} * icf_i\right)^2}} \\ &= \frac{f_{ij} * icf_i}{\sqrt{\sum_i \left(f_{ij} * icf_i\right)^2}} \end{aligned}$$

As neither icf_i or $\sum_i \left(f_{ij} * icf_i\right)^2$ depend on $\sum_i f_{ij}$, neither does e'_{ij}. □

More generally, any cell-wise L^p-normalization removes this dependency.

Proof.

$$\begin{aligned} e'_{ij} &= \frac{e_{ij}}{\sqrt[p]{\sum_i e_{ij}^p}} \\ &= \frac{f_{ij} * icf_i / \sum_i f_{ij}}{\sqrt[p]{\sum_i \left(f_{ij} * icf_i / \sum_i f_{ij}\right)^p}} \\ &= \frac{f_{ij} * icf_i / \sum_i f_{ij}}{\left(\sum_i f_{ij}\right)^{-1} \sqrt[p]{\sum_i \left(f_{ij} * icf_i\right)^p}} \\ &= \frac{f_{ij} * icf_i}{\sqrt[p]{\sum_i \left(f_{ij} * icf_i\right)^p}} \end{aligned}$$

□

Appendix B

Supplemental Figures and Tables

Table B.1: Lists of non-immune ligands for which G13D was significantly different by Mann-Whitney U post-hoc tests after Kruskal-Wallis H between all models

All Benjamini-Hochberg multiple hypothesis corrected

Italics are non-significant

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Plat	0.0	1.5e-262	1.2e-92	1.7e-169
B2m	0.0	0.0	0.0	0.0
C1qb	0.0	0.0	9.9e-66	2.2e-118
Ptn	0.0	0.0	1.3e-145	1.9e-262
Vim	0.0	0.0	2.5e-281	0.0
Lcn2	0.0	0.0	2.6e-37	6.7e-213
Fn1	0.0	0.0	2.4e-134	2.0e-271
Apoe	0.0	0.0	0.0	0.0
Mmp7	0.0	0.0	0.0	0.0
C1qa	2.8e-308	0.0	7.7e-65	7.0e-64
Ccl4	9.1e-278	2.1e-36	1.2e-74	*1.1e-01*

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Cxcl2	7.3e-266	9.2e-80	4.1e-49	5.9e-02
S100a9	7.9e-258	2.3e-39	1.8e-62	4.9e-05
S100a8	2.3e-243	1.5e-28	1.7e-67	*2.5e-01*
Igfbp4	9.7e-238	3.9e-138	3.4e-73	4.6e-190
Lama5	7.5e-236	2.9e-156	7.4e-60	4.5e-121
Rbp4	9.2e-217	4.9e-15	2.2e-76	3.3e-06
Col1a2	2.4e-212	2.4e-18	5.1e-94	2.3e-02
Psap	1.6e-199	7.8e-164	1.0e-15	1.2e-45
Sema3f	3.4e-192	4.0e-133	8.8e-47	1.2e-93
Agrn	9.0e-173	6.4e-164	5.1e-50	2.6e-74
Calr	4.7e-171	1.6e-155	1.4e-52	3.5e-82
Lgals3bp	4.0e-168	6.5e-157	9.4e-61	5.2e-92
Thbs1	4.7e-159	2.1e-124	*3.4e-01*	1.9e-08
Col3a1	5.2e-158	8.2e-20	4.9e-74	2.1e-03
Ccl3	1.1e-154	3.2e-41	3.6e-25	8.9e-04
Mfge8	5.2e-148	2.5e-134	1.1e-46	2.4e-75
Sema5a	1.5e-146	8.7e-29	3.8e-136	7.6e-63
Il1rn	8.9e-146	2.9e-49	5.6e-109	4.1e-07
Sema4g	4.7e-138	1.8e-39	8.9e-136	6.6e-37
Nlgn2	6.9e-133	1.2e-86	7.4e-30	8.6e-67
Cubn	1.4e-132	1.7e-72	2.8e-31	7.6e-66
Cdh1	2.5e-119	5.5e-40	3.9e-10	1.3e-06
Col1a1	1.6e-117	3.3e-10	6.1e-51	*4.0e-01*
Csf1	1.3e-114	1.9e-91	1.3e-21	4.9e-57
Sema3a	4.3e-114	8.6e-78	8.5e-28	1.6e-62

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	p(H) across models	p(U) G13D vs		
		A146T	G12D	WT
Rps19	1.9e-110	7.1e-103	2.1e-85	1.1e-48
Il18	5.9e-108	7.1e-46	3.0e-43	*4.0e-01*
Sema3c	5.5e-105	3.9e-75	2.3e-28	5.0e-83
Col9a3	1.3e-103	1.0e-54	2.1e-24	2.3e-51
Gas6	1.1e-100	2.1e-80	4.2e-87	2.3e-33
Rgmb	7.6e-99	1.6e-05	2.5e-41	5.4e-08
Lama3	3.7e-93	3.5e-05	7.6e-15	1.2e-92
Rtn4	2.3e-88	1.5e-14	1.1e-02	6.6e-71
App	1.7e-87	2.9e-67	2.3e-05	2.3e-07
Lamc2	3.9e-83	1.2e-03	1.1e-04	1.6e-33
L1cam	1.8e-82	5.3e-56	7.6e-75	1.6e-23
Sema4a	8.5e-78	2.1e-13	2.3e-76	1.6e-20
Pgf	6.7e-76	5.0e-60	1.9e-17	4.1e-32
Kitl	7.4e-76	4.1e-67	9.4e-09	2.3e-36
Pdgfc	1.6e-72	3.1e-41	1.5e-18	1.1e-35
Igf1	3.3e-70	1.7e-41	4.0e-17	1.5e-45
Vegfa	4.1e-67	2.3e-36	2.5e-03	3.9e-02
Hp	1.2e-63	1.1e-03	2.6e-20	*3.9e-01*
Efna4	5.3e-63	5.7e-48	7.9e-13	7.1e-36
Hsp90aa1	1.6e-61	6.0e-61	5.6e-33	1.6e-27
Dll4	1.7e-60	1.7e-17	1.3e-10	*2.4e-01*
Il1a	3.5e-53	6.4e-04	2.4e-20	2.2e-07
Adam9	1.2e-52	7.5e-03	1.5e-39	1.1e-02
Cgn	2.2e-52	2.8e-32	5.4e-52	1.7e-14
Cd274	1.0e-51	7.9e-52	1.0e-02	5.9e-14

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Hbegf	3.7e-51	3.2e-23	4.3e-02	*2.5e-01*
Dlk1	2.4e-50	1.2e-56	3.1e-08	2.3e-09
Tnf	8.5e-50	4.4e-37	*2.9e-01*	3.2e-18
Sema3e	1.6e-48	1.4e-37	6.6e-11	6.1e-24
Cd14	1.0e-47	1.5e-03	1.4e-36	*2.5e-01*
Dcn	4.1e-47	2.1e-09	3.7e-23	*3.9e-01*
Il7	4.1e-47	1.6e-31	2.9e-09	4.0e-02
Serping1	1.8e-45	3.7e-30	1.5e-10	3.4e-34
Lin7c	4.7e-45	1.1e-30	1.9e-40	5.4e-10
Cxcl16	6.2e-45	8.1e-38	2.3e-04	1.7e-24
Sema7a	2.2e-43	8.7e-23	2.0e-18	8.0e-39
Efna1	2.8e-43	1.2e-25	2.6e-02	2.2e-03
Lamb2	4.6e-43	8.4e-26	2.9e-10	6.2e-31
Vegfb	2.2e-41	6.1e-16	3.4e-05	*1.1e-01*
Dll1	5.4e-38	2.0e-06	2.4e-10	1.1e-04
Il2	2.8e-37	2.3e-23	4.4e-09	3.6e-33
Col4a1	5.5e-37	2.4e-21	4.3e-04	9.8e-31
Ccl28	5.9e-37	5.9e-07	1.8e-26	6.4e-29
Liph	8.2e-37	6.0e-07	4.1e-33	3.2e-04
Tgfb3	7.6e-36	1.5e-25	1.7e-08	3.4e-19
Pdgfd	3.2e-34	1.3e-21	7.4e-09	2.7e-21
Ntn4	2.1e-33	4.7e-33	*2.4e-01*	8.3e-08
Gpi1	7.0e-33	9.9e-03	6.5e-18	*1.6e-01*
Mmp9	2.4e-32	2.0e-09	4.5e-05	3.7e-03
Adam12	4.6e-32	3.7e-20	6.1e-07	2.8e-21

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Mmp13	7.2e-32	4.3e-06	*2.3e-01*	1.0e-08
Pdgfb	7.6e-32	8.2e-25	*2.6e-01*	6.4e-13
Vcam1	9.8e-32	1.2e-26	1.3e-08	7.7e-16
Agt	7.9e-31	5.1e-12	4.7e-04	*6.0e-02*
Cxcl13	1.0e-30	5.3e-25	*1.2e-01*	2.4e-14
Lamb1	1.6e-30	1.9e-27	6.8e-05	1.3e-16
Cxcl10	1.2e-29	4.5e-29	4.0e-06	2.6e-12
Adam10	1.5e-29	2.0e-07	*1.0e-01*	3.2e-12
Cx3cl1	2.5e-29	*1.9e-01*	5.4e-16	3.6e-02
Spint1	3.4e-29	2.3e-02	2.1e-07	3.7e-05
Col4a2	2.5e-28	2.6e-15	*3.4e-01*	4.8e-18
Bmp2	6.4e-28	*3.7e-01*	6.8e-17	3.3e-03
Efna5	1.2e-27	2.6e-15	2.0e-07	8.7e-19
Sema6d	3.6e-26	3.9e-19	2.4e-09	3.1e-02
Sema3b	3.1e-25	6.4e-03	1.6e-20	*1.1e-01*
Sorbs1	8.9e-25	5.6e-07	1.8e-04	1.6e-02
Col14a1	1.5e-24	2.3e-16	3.4e-03	2.4e-18
Ctf1	2.0e-24	4.1e-18	*1.6e-01*	9.3e-14
Egf	2.3e-24	7.8e-04	3.7e-07	8.5e-05
Ptdss1	3.4e-24	4.8e-06	5.0e-24	2.4e-03
Adam15	1.1e-22	1.9e-02	1.3e-09	*2.6e-01*
Col18a1	3.8e-22	6.2e-16	*1.4e-01*	1.6e-13
Jag1	4.2e-22	6.7e-03	1.2e-06	*6.0e-02*
Adam17	9.4e-22	2.0e-04	1.1e-05	*7.3e-02*
Bst1	1.6e-21	4.3e-13	*2.7e-01*	8.3e-06

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	p(H) across models	p(U) G13D vs		
		A146T	G12D	WT
Edn1	2.5e-21	3.0e-13	*4.1e-01*	9.1e-08
Prss23	5.6e-21	2.4e-16	*3.9e-01*	2.7e-08
Slit2	9.7e-21	6.0e-18	4.5e-03	2.7e-12
Lrpap1	2.5e-20	1.1e-05	2.4e-04	*7.5e-02*
Il10	2.4e-19	*4.3e-01*	1.1e-06	*2.5e-01*
Ccl20	5.2e-19	*4.5e-01*	8.7e-12	3.9e-06
Mdk	5.3e-19	3.5e-09	1.5e-07	9.2e-19
C3	1.1e-18	8.9e-14	*2.8e-01*	5.3e-10
Il34	3.1e-18	9.1e-09	2.4e-02	*2.6e-01*
Bmp3	9.8e-18	9.4e-03	1.1e-13	5.7e-02
Sema4d	1.3e-17	7.4e-14	4.2e-02	6.4e-12
Serpine1	2.1e-17	3.1e-05	1.1e-02	2.8e-03
Timp1	2.6e-17	6.2e-13	*1.3e-01*	3.3e-11
Il15	3.5e-17	2.5e-08	*1.8e-01*	8.7e-15
Sema4b	1.7e-16	1.3e-06	*9.5e-02*	5.3e-05
Fbln1	2.1e-16	3.9e-15	1.3e-03	5.8e-11
Fgf1	2.7e-16	*2.6e-01*	1.3e-12	*6.1e-02*
Pros1	2.7e-16	7.4e-05	3.1e-02	2.3e-17
Ccl25	9.0e-16	1.5e-07	*3.1e-01*	9.6e-07
Jag2	9.5e-16	6.6e-14	4.4e-03	3.3e-10
Psen1	7.8e-15	*5.6e-02*	2.9e-13	*9.4e-02*
Fgf9	1.2e-14	5.2e-12	*1.3e-01*	4.8e-09
Chad	1.5e-14	4.9e-02	3.1e-13	7.5e-04
Nrtn	1.8e-14	5.8e-12	3.3e-03	1.3e-08
Tfpi	5.8e-13	1.9e-10	*3.3e-01*	*1.8e-01*

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Fbn1	8.8e-13	1.9e-02	6.6e-09	1.6e-02
Nmb	9.0e-13	1.9e-09	1.6e-03	3.3e-11
H2-M3	4.5e-12	6.1e-04	3.2e-03	*4.3e-01*
Il6	8.2e-11	*3.2e-01*	8.3e-03	1.3e-06
Mmp2	2.3e-10	*6.2e-02*	3.7e-02	1.2e-03
Tnfsf8	2.8e-10	2.4e-10	3.6e-04	5.0e-06
Itih2	3.0e-10	*3.9e-01*	2.6e-05	2.6e-04
Ly86	2.0e-09	6.0e-10	*1.5e-01*	*8.2e-02*
Nrg3	2.3e-09	1.1e-02	*3.1e-01*	5.9e-06
Hspg2	4.5e-09	2.3e-05	*1.5e-01*	*3.0e-01*
Bgn	1.1e-08	1.3e-03	8.7e-06	*4.1e-01*
Col8a1	1.3e-08	3.3e-07	*3.6e-01*	*7.3e-02*
Hras	2.1e-08	7.2e-09	*1.4e-01*	3.0e-04
Ereg	2.7e-08	*3.7e-01*	9.9e-03	*6.2e-02*
Efnb1	3.4e-08	4.0e-02	1.3e-06	*4.7e-01*
Cfh	7.9e-08	*1.2e-01*	*7.1e-02*	1.0e-02
Tgfb1	1.1e-07	3.8e-08	*2.9e-01*	2.2e-03
Ccl24	2.2e-07	1.4e-02	8.5e-05	6.1e-07
Col5a2	2.3e-07	1.1e-02	1.5e-05	*2.7e-01*
Btc	3.1e-07	3.5e-04	*3.3e-01*	1.4e-06
Efna2	7.8e-07	2.1e-05	3.8e-03	6.2e-06
Il16	1.1e-06	4.4e-03	1.4e-02	2.9e-08
Nid1	1.6e-06	1.5e-02	3.9e-03	*2.8e-01*
Adam2	1.9e-06	3.4e-07	3.0e-04	9.8e-04
Vcan	3.0e-06	*4.8e-01*	4.2e-04	3.3e-03

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Efemp2	3.8e-06	1.3e-05	*1.1e-01*	2.7e-05
Col4a3	4.1e-06	1.8e-06	*6.7e-02*	2.7e-04
Clcf1	4.2e-06	3.0e-03	*8.8e-02*	2.4e-07
Fbln2	8.9e-06	*9.5e-02*	3.1e-04	*4.0e-01*
Ltf	1.1e-05	3.5e-02	*1.1e-01*	*9.9e-02*
Vwf	3.1e-05	*3.2e-01*	2.4e-02	*1.8e-01*
Cxcl12	4.8e-05	*4.1e-01*	9.8e-03	*2.5e-01*
Itgb3bp	5.8e-05	*3.5e-01*	5.8e-03	*4.1e-01*
Efnb2	1.5e-04	*1.7e-01*	2.3e-05	4.8e-02
Edil3	1.9e-04	1.5e-04	*6.9e-02*	4.5e-04
Flt3l	3.3e-04	3.9e-03	*2.4e-01*	9.4e-05
Plau	3.5e-04	1.3e-05	*1.2e-01*	2.9e-02
Cd28	4.4e-04	5.8e-05	*2.4e-01*	1.2e-02
Col5a3	8.4e-04	1.3e-03	*4.1e-01*	7.7e-03
Icam1	1.3e-03	1.7e-02	*4.5e-01*	8.8e-03
Serpinc1	2.9e-03	*2.2e-01*	3.8e-02	2.5e-04
Mfng	3.0e-03	1.3e-02	*1.2e-01*	4.0e-04
Timp3	3.4e-03	*2.3e-01*	*4.0e-01*	4.2e-04
F7	4.4e-03	1.0e-04	*2.7e-01*	*1.8e-01*
Bmp4	4.5e-03	8.5e-03	*3.2e-01*	2.5e-03
Col4a5	4.8e-03	*6.5e-02*	*3.7e-01*	3.7e-02
Qrfp	9.3e-03	8.1e-03	*8.8e-02*	2.3e-03
Icosl	1.0e-02	8.4e-03	2.6e-03	*1.6e-01*
Efnb3	1.5e-02	4.5e-03	*3.1e-01*	2.0e-02
Mst1	2.2e-02	*1.4e-01*	8.5e-03	*2.4e-01*

Continued on next page

Table B.1 – continued from previous page

Ligand in Non-immune	**p(H) across models**	**p(U) G13D vs**		
		A146T	**G12D**	**WT**
Btla	2.5e-02	*4.3e-01*	*5.4e-02*	1.8e-02
Has2	2.8e-02	3.5e-02	*4.8e-01*	3.8e-02
Ntn1	3.7e-02	*2.0e-01*	*3.4e-01*	3.8e-03
Pf4	4.8e-02	*6.7e-02*	*2.6e-01*	*4.2e-01*

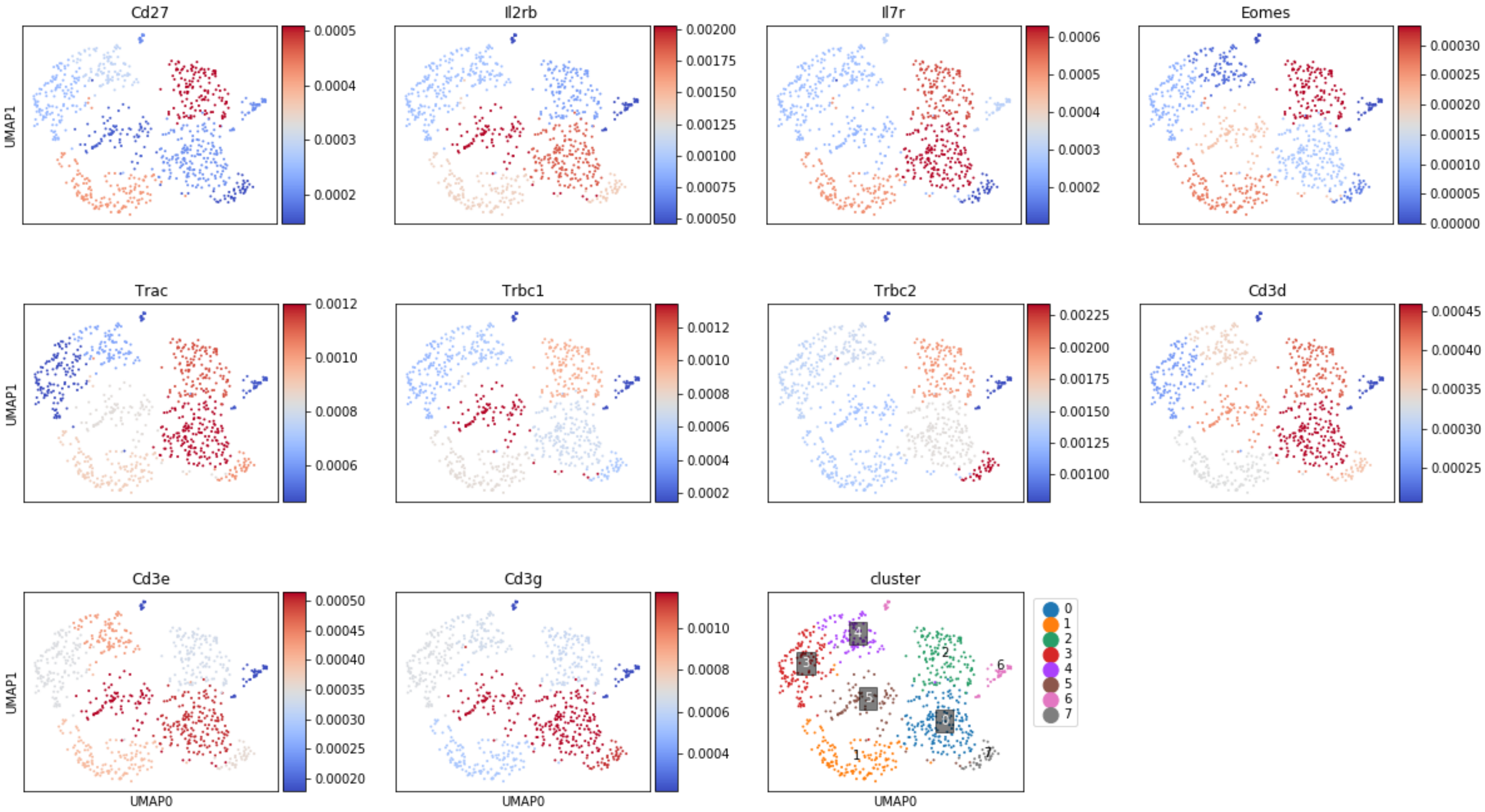

Figure B-1: PCA-UMAP visualization of the T/NK subset with some markers using the first 50 PCs and $\sqrt{n_{cells}}$ neighbors, colored by cluster mean
Top row is NK-cell markers
Second and bottom rows are T-cell markers
The last plot is the Leiden clustering results, with $\gamma = 1$

Bibliography

[1] BB Ancrile, KM O'Hayer, and CM Counter. Oncogenic ras-induced expression of cytokines: a new target of anti-cancer therapeutics. *Molecular Interventions*, 8:22–27, February 2008.

[2] C Blum et al. The expression ratio of map7/b2m is prognostic for survival in patients with stage ii colon cancer. *International Journal of Oncology*, 33:579–584, April 2008.

[3] DM Church et al. Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091, July 2011. doi:10.1371/journal.pbio.1001091; GRCm38.

[4] M Cohen et al. Lung single-cell signaling interaction mapreveals basophil role in macrophage imprinting. *Cell*, 175:1031–1044, November 2008.

[5] M Efremova, M Vento-Tormo, SA Teichmann, and R Vento-Tormo. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, 15:1484–1506, February 2020.

[6] T Er, C Chen, L Bujanda, and M Herreros-Villanueva. Clinical relevance of kras mutations in codon 13: Where are we? *Cancer Letters*, 343:1–5, February 2014.

[7] A Fukuda et al. Stat3 and mmp7 contributeto pancreatic ductal adenocarcinoma initiationand progression. *Cancer Cell*, 19:441–455, April 2011.

[8] G Gambardella and D di Bernardo. A tool for visualization and analysis of single-cell rna-seq data based on text mining. *Frontiers in Genetics*, 10:734, August 2019. doi:10.3389/fgene.2019.00734.

[9] AL Haber, M Biton, N Rogel, et al. A single-cell survey of the small intestinal epithelium. *Nature*, 551:333–339, November 2017. doi:10.1038/nature24489.

[10] KM Haigis. Kras alleles: The devil is in the detail. *Trends in Cancer*, 3:686–697, October 2017.

[11] KB Halpern et al. Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nature Biotechnology*, 36:962–970, September 2018. doi:10.1038/nbt.4231.

[12] CN Heiser and KS Lau. A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. *bioRχiv*, 2019. preprint; doi:10.1101/684340.

[13] CA Herring, A Banerjee, ET McKinley, AJ Simmons, J Ping, JT Roland, JL Franklin, Q Liu, MJ Gerdes, RJ Coffey, and KS Lau. Unsupervised trajectory analysis of single-cell rna-seq and imaging data reveals alternate tuft cell origins in the gut. *Cell Systems*, 6:37–51.e9, January 2018.

[14] B Jacobs et al. Amphiregulin and epiregulin mrna expression in primary tumors predicts outcome in metastatic colorectal cancer treated with cetuximab. *Journal of Clinical Oncology*, 27:5068–5074, October 2009.

[15] D Kim, G Pertea, C Trapnell, H Pimentel, R Kelley, and SL Saltzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, April 2013. doi:10.1186/gb-2013-14-4-r36.

[16] R Kim, Y Suh, K Yoo, Y Cui, Hyeonmi Kim, M Kim, IG Kim, and S Lee. Activation of kras promotes the mesenchymal features of basal-type breast cancer. *Experimental & Molecular Medicine*, 47:e137, January 2015.

[17] AM Klein, L Mazutis, I Akartuna, N Tallapragada, A Veres, V Li, L Peshkin, DA Weitz, and MW Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161:1187–1201, May 2015.

[18] Manu P Kumar. *Computational analysis of cell-cell communication in the tumor microenvironment*. PhD dissertation, Massachussets Institute of Technology, Department of Biological Engineering, May 2019.

[19] MP Kumar, J Du, G Lagoudas, Y Jiao, A Sawyer, DC Drummond, DA Lauffenburger, and A Raue. Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics. *Cell Reports*, 25(6):1458–1468, November 2018. doi:10.1016/j.celrep.2018.10.047.

[20] B Langmead and SL Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–359, March 2012. doi:10.1038/nmeth.1923; index for mm10.

[21] P Liu, Y Wang, and X Li. Targeting the untargetable kras in cancer therapy. *Acta Pharmaceutica Sinica B*, 9:871–879, September 2019.

[22] Q Liu, CA Herring, Q Sheng, J Ping, AJ Simmons, B Chen, A Banerjee, W Li, G Gu, RJ Coffey, Y Shyr, and KS Lau. Quantitative assessment of cell population diversity in single-cell landscapes. *PLoS Biology*, 16:e2006687, October 2018.

[23] A Lièvre et al. Kras mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Research*, 66:3992–3995, April 2006.

[24] L McInnes, J Healy, and J Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, February 2018. arXiv:1802.03426v2.

[25] NIH NCI. About the ras initiative. online, September 2016. cancer.gov/research/key-initiatives/ras/about.

[26] V Petukhov, J Guo, N Baryawno, N Severe, DT Scadden, MG Samsonova, and PV Kharchenko. dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell rna-seq experiments. *Genome Biology*, 19:78, June 2018. doi:10.1186/s13059-018-1449-6.

[27] IA Prior, PD Lewis, and C Mattos. A comprehensive survey of ras mutations in cancer. *Cancer Research*, 72:2457–2467, May 2012.

[28] CJA Punt, M Koopman, and L Vermeulen. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nature Reviews Clinical Oncology*, 14:235–246, April 2017.

[29] JA Ramilowski et al. A draft network of ligand-receptor-mediated multicellular signaling in human. *Nature Communications*, 6:7866–7877, July 2015. doi:10.10138/ncomms8866.

[30] P Rawla, T Sunkara, and A Barsouk. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad Gastroenterologiczny*, 14:89–103, January 2019.

[31] Russo et al. Mutational analysis and clinical correlation of metastatic colorectal cancer. *Cancer*, 120:1482–1490, May 2014. doi:10.1002/cncr.28599.

[32] G Smith, FA Carey, J Beattie, MJV Wilkie, TJ Lightfoot, J Coxhead, RC Garner, RJC Steele, and C Wolf. Mutations in apc, kirsten-ras, and p53—alternativegenetic pathways to colorectal cancer. *PNAS*, 99:9433–9438, July 2002.

[33] B Stolze, S Reinhart, L Bulllinger, S Fröhling, and C Scholl. Comparative analysis of kras codon 12, 13, 18, 61 and 117 mutations using human mcf10a isogenic cell lines. *Scoentific Reports*, 5:8535, February 2015. doi:10.1038/srep08535.

[34] VA Traag, L Waltman, and NJ van Eck. From louvain to leiden: guaranteeing well-connected communities. *arXiv*, October 2019. arXiv:1810.08473v3.

[35] M Uhlén et al. Tissue-based map of the human proteome. *Science*, 347:394–404, January 2015. doi:10.1126/science.1260419.

[36] Y Wang, R Wang, S Zhang, S Song, C Jiang, G Han, M Wang, J Ajani, A Futreal, and L Wang. italk: an r package to characterize and illustrate intercellular communication. *bioRχiv*, online, January 2019.

[37] FA Wolf, P Angerer, and FJ Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15–20, February 2018. doi:10.1186/s13059-017-1382-0.

[38] DR Zerbino et al. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, November 2017. doi:10.1093/nar/gkx1098; GTF for mm10; Retrieved Nov 2019.