The repercussions of recognition: imprints of T cells on the tumor microenvironment

Slagter, M.

CHAPTER 6

Discussion

In this thesis, I present a number of studies of cytotoxic T cells in the tumor microenvironment. This work demonstrates that for a multitude of tumors, the landscape of immunologically foreign neoantigens is substantial (**Chapter 2**), allowing for the application of immunotherapy beyond its poster child, melanoma. Following up on this, we evaluated T cell checkpoint blockade therapy in triple negative breast cancer and identified early molecular markers of clinical response (**Chapter 3**), further supporting the notion that T cell attack of varying tumor types is realistic. Moving over to more fundamental research projects on how T cell activity sculpts the tumor microenvironment, we first demonstrated that when T cells encounter cells that present the 'right' (cognate) antigen in the tumor microenvironment, they release IFN-γ that reaches (far) beyond the antigen-presenting cell, whereas TNF-α is strongly localized (**Chapter 5**). We also tested for the loss of neoantigenic mutations as a consequence of T cell pressure in treatment-naïve tumors, but found no evidence for it (**Chapter 4**). A high level summary of these studies would be that T cell recognition of cancer is expected to occur for many types of cancer (**Chapter 2**), can be boosted using combinations of T cell checkpoint blockade and other therapies in TNBC (**Chapter 3**) and that its repercussions reach far and wide in the tumor microenvironment (**Chapter 5**) but do not occur in all matters one would have expected them to (**Chapter 4**). To conclude my thesis, I want to delve deeper into algorithmic requirements to enable highly multiplexed transcriptome based stimulus inference, with the aim of enabling the simultaneous study of many (T cell-secreted) cytokines in the tumor microenvironment.

Cytokines mediate communication between cells of the immune system and other cell types and vice versa, and thereby form key regulators of the immune response. Thus, our understanding of overall tumor biology and immunology would benefit from the ability to study the dissemination of cytokines and other stimuli with high spatiotemporal resolution. While previous work has shed light on the spatial effects of single or a small number of cytokines simultaneously, a lack of experimental tools for multiplexed investigation has impeded a comprehensive understanding of the spatiotemporal spreading of the many cytokines that collectively mediate and orchestrate (tumor) immunity.

As cytokines effectuate transcriptional changes in the cells that encounter them, a promising way to study their *functional* dissemination (i.e., the reach of their gene regulatory capacity) is to infer it from transcriptional read-outs. This leverages the reporting capability that nature has already granted us, and frees us from having to engineer in reporters ourselves. Given the large diversity in response kinetics among such endogenous reporter genes, we might not be limited to only inferring the nature of the encountered stimuli. Possibly, we could also infer the duration and/or concentration of exposure. Progress in the study of cell communication in general and cytokine dissemination in particular will likely benefit from having available a toolkit for stimulus inference of such nuance.

Here, I will first delve deeper into considerations regarding application of transcription based stimulus inference, that we did not fully explore in **Chapter 5**. Next, I'll discuss ideas for algorithms to deal with the obstacles in this approach. By summarizing the understanding I have built up around this topic, I hope to provide guidance to those interested in building upon our work.

I will start with an overview of the procedure of applying transcriptional/RNA-based stimulus inference (Figure 6.1). First, one would decide whether a specific cell type(s) will serve as *reporter(s)* (**Chapter 5**) or whether a mixture of cell types will be used[1]. The reporting cell type could be engrafted in a model system, typically a recipient mouse, but for certain questions an organoid cell culture could also suffice. Alternatively, endogenous cells can also be used, obviating the need for an engraftment step. Next, the experimental model would be exposed to a manipulation of interest (e.g. activation of T cells or other immune cells). Then, the reporter cells would be harvested and can potentially be separated from non-reporter cells using e.g., flow cytometry assisted cell sorting (FACS), which can be facilitated by transducing the reporter cells with a fluorescent reporter before engrafting them in the model system. Cells would then be extracted and sequenced using (single cell) RNA-seq protocols, thereby yielding the *query* data set on which stimulus inference will subsequently be performed. In parallel, one would have to train models of gene expression in the reporting cell type, and possibly also cell viability, in response to cytokines and other stimuli of interest. This can be done using a compendium of *reference* RNA-seq samples to learn from, ideally of the exact same cell type as that of the reporter cell type as different cell types can differ tremendously in their transcriptional response to cytokines[1]. To increase power and reduce cost, this reference dataset may be obtained using bulk RNA-seq and is composed of samples that were exposed to the various stimuli of interest. Finally, the analyst would infer the stimulus exposure of the query (single cell RNA-seq) expression data using the cytokine response models trained on the (bulk RNA-seq) reference dataset.

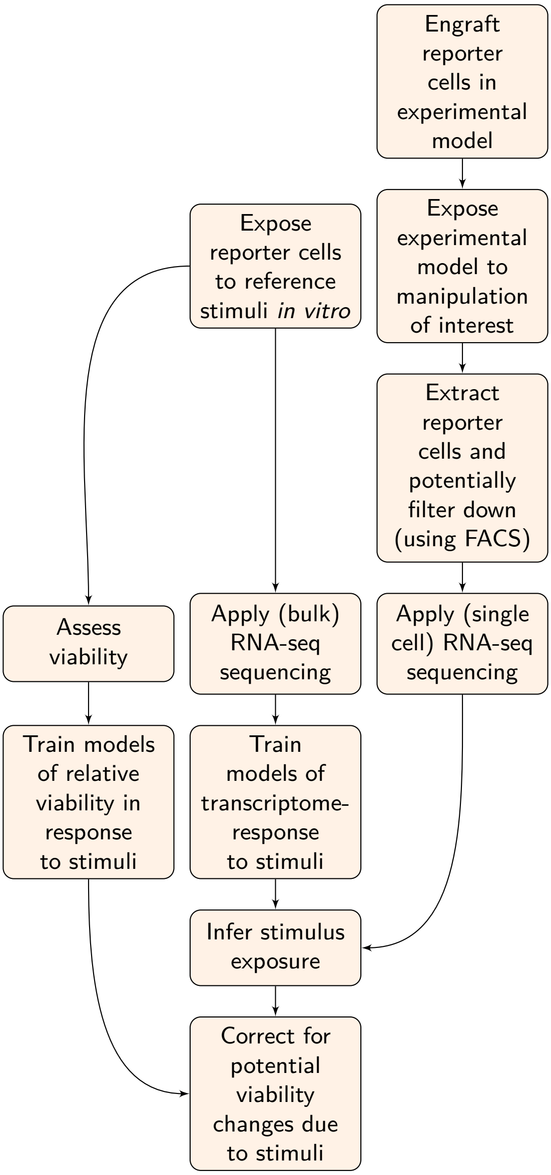
Engraft reporter cells in experimental model
Expose experimental model to manipulation of interest
Expose reporter cells to reference stimuli *in vitro*
Extract reporter cells and potentially filter down (using FACS)
Assess viability
Apply (bulk) RNA-seq sequencing
Apply (single cell) RNA-seq sequencing
Train models of relative viability in response to stimuli
Train models of transcriptome-response to stimuli
Infer stimulus exposure
Correct for potential viability changes due to stimuli

Figure 6.1: An exemplary general process overview for stimulus inference using transcriptomic read-outs

This approach has clear attraction and potential, but there are also a number of potential obstacles to be mindful of in applying it. I will first detail these considerations and then discuss analytical approaches to stimulus inference.

(i) overlap in reporter genes between studied and also ignored stimuli, complicating discrimination between stimuli.

(ii) (differing) sources of confounding variation in the query and reference data.

(iii) modulation of the reporter cell pool's size, by e.g. inducing differentiation into other cell types, cell state transitions, reduced viability or cell death.

(i) Overlap in reporter genes between studied signals

A conceptually simple approach for stimulus inference from (single cell) transcriptome data, is to identify so-called *mono-responsive* genes, which show convincing response to only one of the studied stimuli and minimal to no response to any other stimuli, and sum them for the query transcriptomes. This *gene set* approach should work especially well if the studied cytokines show a high degree of uniqueness, i.e., non-overlap, in terms of their reporter genes. However, the signal transduction pathways that couple stimuli sensing receptors to gene regulatory mechanisms frequently typically display a high degree of crosstalk, causing reporter genes to be shared between stimuli. This can limit the applicability of this approach in two ways. First, the reporter genes may overlap with those of other stimuli in the *in vivo* system of application, at risk of causing false positive inferences if this is not properly accounted for in the experimental design. In **Chapter 5**, the unstimulated (control) samples we collected in addition to the T cell-exposed (test) samples allowed us to establish the baseline level of reporter gene expression signal in absence of the experimental manipulation. Second, overlap between the stimuli of interest will limit our capability of differentiating between them. With a growing number of cytokines or other stimuli, unique reporter genes for the cytokines of interest will become increasingly rare. Supporting this notion, a high degree of overlap in transcriptional response to various cytokines has been found using a large aggregate of publicly available transcriptome datasets[2]. Similarly, related cytokines were shown to overlap in terms of activated genes in single cell data[1]. For the murine system we studied in the last part of Chapter 5, mono-responsive genes were scarce to absent. Here, all genes that strongly respond to TGF-β turned out to negatively respond to IFN-γ. Clearly, analytical methodology to deal with this convolvement of factors is a first requirement to studying large numbers of stimuli, especially when these stimuli display a large degree of reporter overlap or when one aims to discriminate between different concentrations and/or durations of the stimuli. In addition to the challenges reporter sharing presents, one clear advantage is that absence of expression of such clusters allows for the inference of absence of not just one but all overlapping stimuli.

(ii) Interference of the transcriptome by confounding sources of variation

A second complication is that the reporter genes, selected for their stimulus reporting behavior, can be subject to modulation by confounding technical factors and/or other biological processes.

A first technical problem may stem from differences in measurement (platform) technology between the (unlabelled) query data and the labelled reference data. Considering the large number of stimuli experimentalists will want to screen to generate reference datasets, bulk RNA-seq is attractive because of its cost effectiveness. However, when using these bulk reference data in conjunction with query single cell data, one will have to deal with the ensuing technical variation between the two platforms. Bulk RNA-seq data typically features millions of reads and detects 10s of thousands of genes per sample, typically showing a rather homogeneous distribution of read counts across genes. In contrast, (droplet-based) single cell RNA-seq only detects a few thousand genes per single cell, with count distributions that are strongly dominated by a relatively small set of genes. When not properly accounted for, the many zeros in the single-cell data would result in low overall similarity to the bulk RNA-seq data, limiting the utility of query-to-reference similarity metrics.

A second complication in relating single cell query to bulk reference data is the potential activity of cellular processes that will be averaged in bulk reference data but clearly not so in the single cell query data. An example of a such a cell-intrinsic process is the cell cycle, which affects thousands of genes in a periodic fashion[3]. An obvious, experimental remedy for cell-intrinsic sources of variation could, in theory, be to neutralize the interfering process with a drug. However, this risks affecting the transcriptome as well, biasing downstream stimulus inference. A first *analytical* alternative approach is to a priori ’regress out’ the confounding process. All popular single cell RNA-seq analysis software packages offer functionality for a simple, statistical correlation-based version of this. However, as the stimuli of interest could also modulate the confounding process (e.g. IFN-γ inhibits cell cycle progression), this could remove much of the informative value pertaining to the stimulus of interest, throwing away the baby with the bath water. A more sensitive approach to offset cell-intrinsic variation could be to apply a deconvolution step in which active processes are identified in a more data-driven, unsupervised manner, allowing for more granular removal of confounding processes. Recent mathematical advances by Karin et al. (2023) have allowed upfront deconvolution of biological processes in (single cell) expression data through spectral analysis of the associated covariance matrix[4,5]. Rather than relying on pre-identified marker genes of the confounding process to deconvolve the single cell transcriptomes, the authors only assume a ’topology’ of variation in gene expression caused by the confounding process[5]. The topology can roughly be understood as the general shape of the path that cells travel in ’state space’ as they progress through the confounding process. Mapping the cells to the topology allows for post-hoc identification of genes that align with the topology and filtering of the transcriptome for gene expression that is aligned with the topology. For the specific case of the cell cycle, a circular topology was chosen[5]. Future work could use ’gold standard’ single cell datasets, that contain cells exposed to a well-controlled stimulus *in vitro* as reference data, to test whether more complex topologies than the circular one can clean the data of additional confounding processes and improve accuracy on the downstream stimulus inference task.

Third, stimuli or confounding signals that are restricted to either the reference or to the query settings might (partially) overlap in terms of reporter genes with the stimuli of interest. Probably most prone to this are *in vivo* settings (e.g. the query single cell data in our case), wherein it is likely that many biological processes and cell types are active that are absent from the well-controlled *in vitro* reference conditions of a purified cell type. The possibility of such confounding warrants the design of experiments to include unexposed, control, conditions to any 'test' condition. In **Chapter 5** we included *in vivo* controls without T cells to the *in vivo* test conditions (i.e., with T cells) for this reason. As we'll see below, an analytical approach to remedy this would be a form of data integration that selectively removes covariation that is unique to any of the to be integrated datasets.

For the sake of completeness, I lastly also mention that confounding variation may stem from "mundane" factors like varying batches of chemicals, room temperatures, and different people executing the work.

(iii) Modulation of the reporter cell pool's size

Cytokines and other stimuli of interest may cause (reporter) cells to differentiate into other cell types or to attain differing cell states. They may also directly affect the number of the reporter cells by in- or decreasing proliferation and even have cytotoxic effects. Unless the reporter cell pool is defined in a wide (i.e., unspecific) enough fashion (and e.g. cell state transitions cannot cause reporter cells to go 'out of scope'), such modulation would bias direct and unadjusted stimulus abundance estimates. The solution will be to explicitly model the effect of such stimuli on the cell pool, allowing for post-hoc adjustment of the direct abundance estimates.

A first example of a stimulus that can modulate the reporter cell pool size is the cytokine IL-2, which specifically induces T cells (more specifically: cells with the IL-2 receptor) to proliferate[6]. One can then immediately see that a direct read-out of the number T cells that *appeared* to have experienced IL-2 cells would not directly reflect the number that have actually done so - the reporter T cells may have simply inherited IL-2-reporting transcripts from their ancestors. Stimuli may also decrease the cell pool size. Among the cytokines we studied in **Chapter 5**, we found IFN-γ to cause reduced proliferation and be cytotoxic to OVCAR5 cells in particular. Dying cells lose their cell membrane integrity and thereby lose their compatibility with FACS-based protocols. As such, if FACS-based enrichment of reporter cells is part of the experimental setup and cytotoxicity remains unaccounted for, the number of cells exposed to cytotoxic stimuli will appear lower than it actually is.

However, increased proliferation or even cytotoxicity (if not complete), does not have to be a show-stopping problem. To attain unbiased estimates of the number of stimulus-experienced cells, one can correct for the expected change in cell numbers due to a given stimulus. For this, one could use a parametric model of cell growth or death (a so-called viability model), as frequently employed in drug sensitivity screens[7]. If the analyst is additionally willing to assume that the stimulus level remained constant up until the moment of measurement, correcting for cell pool size modulation effects is trivial. First, the viability model would describe viability as a function of stimulus concentration and/or duration with a monotonic, smooth function. Here, viability values larger than 1 would indicate relative cell growth, whereas values smaller than 1 would indicate decline. Next, the number of cells that have experienced the stimulus can be recovered from the direct (biased) estimates. For this, one would multiply the direct estimates by the inverse of the expected change in viability under the stimulus. This concept can be extended to combinations of stimuli as well, given sufficient laboratory resources to combinatorially screen their effects on viability. Alternatively, if one is additionally willing to assume that the cytotoxic effects are not synergistic but rather additive, the cytotoxicity of combinations of stimuli can be modelled by multiplying those of the individual stimuli.

There may be instances in which reference profiles of multiple concentrations and/or exposure durations per stimulus are available and the analyst aims to discriminate and interpolate between these levels. Viability-correction can also be done for such 'continuous' rather than 'discrete' stimulus exposure. Let **s** be a vector with entries denoting the concentrations and exposure durations, along with other relevant indicators, of the various stimuli under study. Analogous to the 'discrete' case described above, the computation of summary statistics (e.g., the fraction of cells that have experienced stimulus **s** for at least 12 hours) would be done in a weighted rather than unweighted fashion. Here, we would use an integral of the form $\int_D c(\mathbf{s})/v(\mathbf{s})d\mathbf{s}$, wherein D is the domain of interest (e.g., the set of all **s** for which IFN-γ-exposure ≥ 12), $c(\mathbf{s})$ is the fraction of cells directly inferred to have experienced stimulus (or the combination of stimuli) **s** and $v(\mathbf{s})$ is the estimated viability with this stimulus. A caveat here is that estimation accuracy will decrease with increased cytotoxicity (or extreme degrees of cell growth), and by extension is not feasible for (combinations of) stimuli that fully wipe out the reporter cells.

Regression based stimulus inference

Mindful of the obstacles to transcriptome-based stimulus inference we just considered, how do we choose a stimulus inference methodology? In **Chapter 5**, we actually applied two different ones. The first 'geneset' approach is methodologically simple but sufficed for the human OVCAR5 model, in which mono-reporters were readily available and the experimental design helped us to further disambiguate cytokines. However, for the murine NMM cell line we studied in the same chapter, mono-reporters were virtually absent, necessitating a more sensitive approach. Such an approach should maximally leverage the differential magnitudes with which reporter genes are modulated between stimuli, or different concentrations and durations of one stimulus. Regression analysis, in which gene expression levels are quantitatively compared between query and reference transcriptomes, could be the ideal foundation for this. In theory, it can leverage expressional nuance to discriminate between related stimuli, or even between different exposure durations/concentrations of the same stimulus. Here, a query transcriptome (e.g., a single cell or aggregate of multiple similar single cells) is modelled as a weighted sum of expression profiles from the reference dataset. The weights assigned to each of the reference profiles are then interpreted as proportional to the similarities of these reference profiles to the query transcriptome.

In using regression analysis to relate query (single cell) trancriptomes to (bulk) references, the resulting regression coefficients are sensitive to the way the input data are batch-normalized and integrated. In **Chapter 5** we employed pseudo-bulkification (i.e., aggregating transcriptionally similar cells to 'meta'-cells, Figure 6.2.2) to first make the single cell data more similar to the bulk reference data. We next employed a computational trick to work around the requirement of (near-)perfect comparability. Specifically, we noticed that the presence of a stimulus can not just be shown by high regression coefficients to the associated reference sample, but can also be inferred from obtaining an increased *reconstruction error* when omitting the corresponding reference sample from the analysis. Comparing the 'full model' error, obtained using the full set of reference samples, by a 'partial' one, obtained using only a subset of the reference profiles (i.e., by assessing the relative error of these two models), allowed us to assess the importance of the left out reference samples and the potentially unique gene expression patterns they contained that were not recoverable from any of the remaining samples. As this relative reconstruction error can be computed for individual query transcriptomes (relating to single cells or single cell neighborhoods), the relevance of individual or groups of reference samples to individual query transcriptomes could be quantified. In the absolute sense, the reconstruction error using the full reference set was quite variable between neighborhoods, probably due to a host of possible reasons (as discussed above). However, assessing the relative reconstruction error inherently corrects for this variability. Such a reconstruction error is more frequently used in regression analysis, typically to assess feature importance. However, In the context of (bulk) RNA-seq deconvolution and stimulus inference, this was to the best of my knowledge a novelty.

With all data being derived of one particular cell line and acquired in one single (our) laboratory, the homogeneous nature of the data in **Chapter 5** permitted us to omit stringent batch effect correction. However, using the relative reconstruction error does not fully obviate the need for sensitive correction of confounding variation. A computational methodology that can appropriately deal with data from heterogeneous sources benefits from being able to draw upon a much broader set of data sets and sources. With this in place, one could compile a compendium of cytokine stimulated transcriptomes from a range of studies[2], maximally leveraging past efforts of the scientific community, and arrive at a more diverse and powerful predictor of cytokine exposure. To leverage such heterogeneous data, one needs to deal with confounding variation 'upfront' and preprocess the data in a manner that filters out (biological or technical) confounding variation. How do we arrive at this point?

A first obstacle is the difference in count distributions between bulk and single cell RNA-seq data (see above), that precludes direct use of preprocessing solutions developed specifically for single cell

RNA-seq data. This is probably why I found that existing single cell RNA-seq integration methodologies, which essentially rely on library size normalization and Euclidean distance-based similarity metrics to relate transcriptomes, did not perform well when applied to the OVCAR5 data from Chapter 5 (data not shown). However, such algorithms did achieve sensible integrations for their intended domain of application, i.e., data sets that were exclusively of single cell nature. If one has access to both bulk and single cell RNA-seq data of the same cell lines and stimuli - like we did - then one viable approach may be to somehow learn how to simulate single cell transcriptomes from bulk RNA-seq samples, such that these simulated cells can then be related to the query single cells. Specifically, one could use a modification of the variational auto-encoder (VAE) for this, which would disentangle the various active transcriptional programs from each other. The VAE would consist of two (deep) neural networks: one encoder network which reduces the dimensionality of bulk reference gene expression to a much smaller number of latent factors, and a second decoder network which reconstructs single cell expression profiles from the latent factors. Each latent factor should represent a (biological) source of variation and captures combinations of correlated non-linear patterns in the data, akin to the principal components in a (linear) principal component analysis. Input to the probabilistic VAE encoder would be the bulk expression data, along with stimulus exposure and batch-related metadata, and the output of the VAE decoder would be the associated single cell transcriptomes. To ensure biological relevance of the identified latent factors, sparsity constraints could be imposed on the VAE's component models, limiting the number of active latent factors per sample. Another approach towards this goal would be to pre-train the VAE on single cell data exclusively. This is less ambitious as it doesn't require the encoder network to additionally learn which processes tend to co-occur in single cells, like it would have to when its input is bulk data. Then, especially the encoder network could be adapted to the task of using bulk data as its source (i.e., an application of transfer learning), or an additional encoder network could be placed in front of the encoder network. Having the VAE in place, one could next simulate tons of single cells for each of the reference samples, including the ones for which no matching ground truth single cells are available. To then finally do stimulus inference for the query *in vivo* data, the frequently used and robust mutual nearest neighbor algorithm[8] could be used to integrate them with the labelled, simulated cells in an unsupervised manner.

Another approach to leveraging data from different labs or lab technologies is to directly relate query and reference transcriptomes in a manner that is robust to distributional differences and confounding variation. In collaboration with Soufiane Mourragui, I developed an approach to integrate data such that it is insensitive to differences in count distributions and that resulted in a good balance between inferential sensitivity and robustness (Data IntegratIon inSensitive To distributional iNbalanCes, DISTINCt, Figure 6.2. It involves relating bulk RNA-seq and single cell RNA-seq data in a higher-order feature space induced by the *Mallow's kernel*[9] combined with Mourragui's domain adaptation algorithm TRANSACT[10]. On our own *in vitro* data, it remained completely insensitive to the distributional differences between the bulk and single cell RNA-seq data, while nearly perfectly inferring IFN-γ stimulus exposure duration (Figure 6).

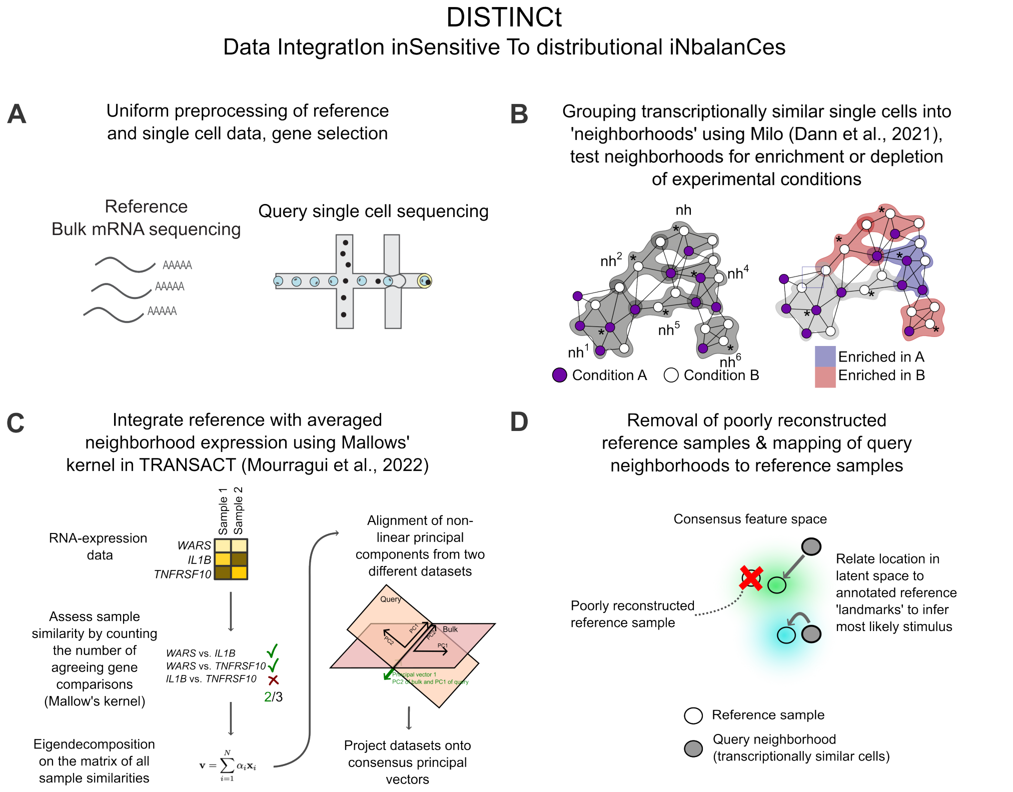
DISTINCt
Data IntegratIon inSensitive To distributional iNbalanCes
A
Uniform preprocessing of reference and single cell data, gene selection
Reference Bulk mRNA sequencing
AAAAA
AAAAA
AAAAA
Query single cell sequencing
B
Grouping transcriptionally similar single cells into 'neighborhoods' using Milo (Dann et al., 2021), test neighborhoods for enrichment or depletion of experimental conditions
nh
nh2
nh4
nh5
nh1
nh6
● Condition A
○ Condition B
Enriched in A
Enriched in B
C
Integrate reference with averaged neighborhood expression using Mallows' kernel in TRANSACT (Mourragui et al., 2022)
RNA-expression data
Sample 1
Sample 2
WARS
IL1B
TNFRSF10
Assess sample similarity by counting the number of agreeing gene comparisons (Mallow's kernel)
WARS vs. IL1B ✓
WARS vs. TNFRSF10 ✓
IL1B vs. TNFRSF10 ✗
2/3
Eigendecomposition on the matrix of all sample similarities
$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$
Alignment of non-linear principal components from two different datasets
Query
Bulk
PC1
PC2
PC1
PC2
Principal vector 1
PC2 of bulk and PC1 of query
Project datasets onto consensus principal vectors
D
Removal of poorly reconstructed reference samples & mapping of query neighborhoods to reference samples
Consensus feature space
Poorly reconstructed reference sample
Relate location in latent space to annotated reference 'landmarks' to infer most likely stimulus
○ Reference sample
● Query neighborhood (transcriptionally similar cells)

Figure 6.2: The Mallow's kernel can be seen as a mathematical function to compute sample similarities, using gene ordering (i.e., all pairwise comparisons between genes) instead of absolute expression levels to relate sample to each other. This makes it robust against the distributional differences between bulk and single cell sequencing samples. From the pairwise sample similarities within any of the individual datasets, one can derive a latent, lower dimensional representation using kernel principal component analysis, a generalization of 'standard' PCA. Oversimplifying, the non-linear principal components from the kernel PCA (with the Mallow's kernel) group comparisons of genes that tend to co-occur. TRANSACT builds on kernel PCA by matching sets of non-linear principal components of the two datasets with each other and interpolating between them to compute so-called consensus features (C). The consensus features capture sources of variation (e.g., gene expression programs) within each dataset that are shared between the two datasets, discarding variation that is unique to any one of the two individual datasets. This makes it robust to, among other confounders, cell-intrinsic sources of covariation (e.g. the cell cycle). It finishes by projecting the reference and query datasets onto the consensus space, allowing direct comparison between the reference and query samples. Proximity in the consensus space can be interpreted as an indication of similar cytokine exposure (D). To further increase inferential strength and speed up computation, one can optionally also a priori aggregate single, transcriptionally similar cells into local averages and thereby denoise RNA expression values, using cell graph based aggregation[11,12] (B).

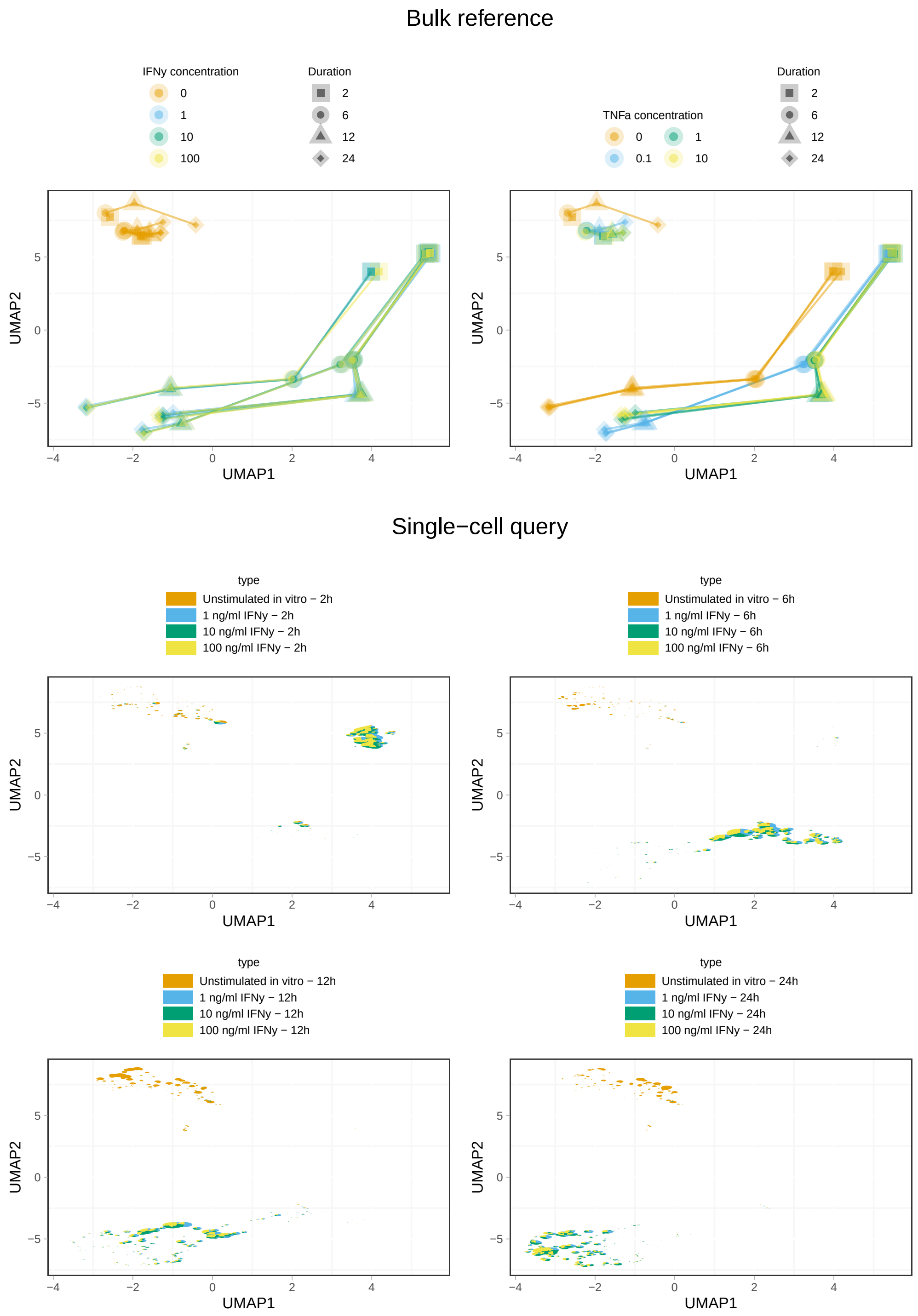
Bulk reference
IFNy concentration
0
1
10
100
Duration
2
6
12
24
TNFa concentration
0
0.1
1
10
Duration
2
6
12
24
UMAP2
UMAP1
5
0
−5
−4
−2
0
2
4
Single−cell query
type
Unstimulated in vitro − 2h
1 ng/ml IFNy − 2h
10 ng/ml IFNy − 2h
100 ng/ml IFNy − 2h
type
Unstimulated in vitro − 6h
1 ng/ml IFNy − 6h
10 ng/ml IFNy − 6h
100 ng/ml IFNy − 6h
type
Unstimulated in vitro − 12h
1 ng/ml IFNy − 12h
10 ng/ml IFNy − 12h
100 ng/ml IFNy − 12h
type
Unstimulated in vitro − 24h
1 ng/ml IFNy − 24h
10 ng/ml IFNy − 24h
100 ng/ml IFNy − 24h

Figure 6.3: Seamless integration of reference bulk and *in vitro* single cell query data sets with DISTINCt. All panels show a UMAP-transformation of samples projected onto the consensus features resulting from TRANSACT-integration in combination with the Mallow's kernel. The top 2 panels show the bulk RNA-seq reference dataset, once colored by IFN-γ-concentration (top row, left) and once by TNF-α-concentration (top row, right). The lines connect samples that were stimulated with the same stimuli and concentrations. Since we controlled the stimulus exposure of these cells experimentally *in vitro*, we can use them for benchmarking purposes, along with unstimulated controls. The bottom 4 panels show the *in vitro* IFN-γ-titration single cell query dataset, with one figure panel per exposure duration. Each circle (pie-chart) in these panels represents one 'neighborhood' of cells, i.e., a variably sized group of transcriptionally similar cells, that is color coded by the exposure concentration of IFN-γ. This single cell dataset shown here consists of all 16 combinations of 4 concentration of IFN-γ (0, 1, 10 and 100 ng/ml) and 4 exposure durations (2, 6, 12 and 24h). The reference reference dataset (top panels) covers these same stimuli but additionally also contains 3 TNF-α concentrations, as well as the combination of IFN-γ and TNF-α. The integrated data permit clear inference of IFN-γ exposure time, but not concentration, for the bulk reference and single cell query data. This can be seen by the clear separation of exposure durations for both the reference and query data. For instance, the first panel (middle row, left) shows the coordinates of cells that were stimulated for 2 hours. Here, unstimulated cells (orange) gravitate towards the top left in UMAP space, where the unstimulated reference samples are also positioned. In contrast, all IFN-γ-stimulated cells are located on the top right, exactly where the bulk reference data of this duration are also positioned. For the other exposure durations, the single cell query and bulk reference data are also well aligned. This shows that the Mallow's kernel represents the input data in an informative manner, capable of discerning different exposure durations of IFN-γ to each other, whereas different concentrations of IFN-γ at any of the evaluated exposure durations are too similar in order to be reliably distinguished. However, this latter unseparability was already present in the untransformed, original data (data not shown). This indicates that gene expression just cannot discriminate between these IFN-γ concentrations. Also, notice how the TNF-α-stimulated cells in the reference are positioned very close to the unstimulated samples. Due to the absence of TNF-α-stimulation in the single-cell query data, TRANSACT mostly discards TNF-α-related variation in the bulk RNA-seq data (*'variational collapse'*).

Currently still open is the question on how to deal with *'variational collapse'* in DISTINCt. DISTINCt finds consensus features using shared variability, and variability that is unique to any one of the datasets will hence - by design - be discarded. This has the consequence that samples that have been stimulated in a manner that uniquely occurs in just one of the two datasets to be integrated will appear to not or barely have sensed this stimulus in the integrated data representation. Variational collapse is observable in Figure 6, wherein the reference samples stimulated with both IFN-γ and TNF-α appear eerily close to the reference samples that were exposed to IFN-γ alone. This is due to the absence of TNF-α-stimulated samples in the query dataset, such that all gene comparisons that are informative in that regard are discarded in the consensus space. One way to (automatically) diagnose this problem would be to assess how accurately the consensus space representation resembles the original input representation (i.e., the reconstruction error), but the mathematics to support this operation have not been developed yet. To get a proxy for the reconstruction error, the relative distance between samples in the original and consensus representations can be considered. For instance, to quantify the reconstructability of sample j, one could consider its distance (i.e., the l^2-norm) to an unstimulated 'anchor' sample in the consensus representation, and divide that by the same sample ratio in the original/untransformed representation. Samples for which this reconstructability value lies near 2 could be deemed as equally well represented in both the original and the integrated representations. In contrast, values closer to 0 would indicate the imprint of a stimulus or process that was not represented in the consensus space, presumably because it wasn't present in the other dataset. To offset such variational collapse, the last query to reference mapping step of the algorithm would weight reference samples based on their reconstruction error, prioritizing samples with high reconstructability, or fully remove (reference) samples for which the reconstruction does not meet some predefined criterion (Figure 6.2.D).

Additional inferential value from alternative splicing information

Isoform specific mRNA quantification, along with large numbers of samples as are typically screened in single cell RNA-seq experiments, allow for estimation of so called RNA velocity[13]. RNA velocity is a time derivative of gene expression which can be used to order cells along temporal and cell differential axes. It might also offer a rich layer of information regarding the timing of cytokine exposure, complementing that of the 'snapshot' RNA abundance-derived estimates that are discussed above and in **Chapter 5**.

In exploratory analyses with our bulk RNA-seq data, which was not acquired using unique molecular identifiers (UMIs) as is typically done in single cell RNA-seq, we obtained noisy estimates of RNA velocity that did not offer informative value in addition to the readily available splicing-agnostic gene expression estimates. However, in the single cell RNA-seq data we acquired using the 10x protocol and with each identified molecule tagged with a UMI, RNA velocity should be much more robustly inferrable. Even though the per sample sequencing depth in single cell RNA-seq is much lower, the large number of samples in conjunction with population based estimates[13] in these datasets offer an opportunity towards increasing the signal-to-noise ratio.

To then properly incorporate RNA velocity in the frameworks that are discussed above, one could set up a reference compendium with single cell RNA-seq. Compiling a reference dataset of single cell data would be more costly but would also give more insight in terms of cell-to-cell response heterogeneity and extrinsic sources of variation. An exciting newer technology to use is VASA-seq, which unlike more traditional single cell-seq protocols is not biased towards the 3' ends of RNA molecules and allows for more accurate RNA velocity estimates[14]. VASA-seq is also applicable to bulk RNA-seq sequencing (personal communication with Soufiane Mourragui). I expect RNA velocity features to especially become useful when deconvoluting transcriptomes affected by many cytokines of overlapping transcriptional signatures. On a related note, multimodal, rather than unimodal, data, for instance additionally employing CITE-Seq[15] with antibodies directed against cell surface proteins relevant to the stimulus in question, could also allow for further increases in inferential accuracy.

Outlook of transcriptome-based cytokine inference

The work my colleagues and I have presented in Chapter 5 forms a useful step in our understanding of cytokine mediated communication between T cells and the tumor microenvironment. I foresee that in the decades to come, the field will continue to generate research tools to study signaling with increasingly high throughput. This increasing throughput is likely to be valuable, as the complex underpinnings of the tumor microenvironment and other biological systems can only be fully elucidated by assessing these dynamic entities in parallel. The increasingly widespread use of spatial single RNA-seq, and its ability to study cells in their physiological context[16], will unlock another crucial layer of information with regard to cytokine dissemination and amplification. In this, deconvolution of the effects of the many cytokines that may be at play will be especially important. As transcriptome based inference has high discriminatory potential and does not require upfront modification of the reporter cell, it's bound to continue to be a potent avenue towards studying stimulus dissemination and cellular crosstalk in biological systems.

References

1. Cui, A. *et al.* Dictionary of Immune Responses to Cytokines at Single-Cell Resolution. *Nature,* 1–8. issn: 1476-4687. (2023) (Dec. 2023).

2. Jiang, P. *et al.* Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nature Medicine* **24,** 1550–1558. issn: 1546-170X (Oct. 2018).

3. Dominguez, D. *et al.* A High-Resolution Transcriptome Map of Cell Cycle Reveals Novel Connections between Periodic Genes and Cancer. *Cell Research* **26,** 946–962. issn: 1748-7838. (2024) (Aug. 2016).

4. Nitzan, M. & Brenner, M. P. Revealing Lineage-Related Signals in Single-Cell Gene Expression Using Random Matrix Theory. *Proceedings of the National Academy of Sciences* **118,** e1913931118. (2024) (Mar. 2021).

5. Karin, J., Bornfeld, Y. & Nitzan, M. scPrisma Infers, Filters and Enhances Topological Signals in Single-Cell Data Using Spectral Template Matching. *Nature Biotechnology* **41,** 1645–1654. issn: 1546-1696. (2024) (Nov. 2023).

6. Ross, S. H. & Cantrell, D. A. Signaling and Function of Interleukin-2 in T Lymphocytes. *Annual review of immunology* **36,** 411–433. issn: 0732-0582. (2024) (Apr. 2018).

7. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell.* issn: 10974172 (2016).

8. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch Effects in Single-Cell RNA-sequencing Data Are Corrected by Matching Mutual Nearest Neighbors. *Nature Biotechnology* **36,** 421–427. issn: 1546-1696. (2024) (May 2018).

9. Jiao, Y. & Vert, J.-P. The Kendall and Mallows Kernels for Permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40,** 1755–1769. issn: 0162-8828, 2160-9292. (2023) (July 2018).

10. Mourragui, S. M. *et al.* Predicting Patient Response with Models Trained on Cell Lines and Patient-Derived Xenografts by Nonlinear Transfer Learning. *Proceedings of the National Academy of Sciences of the United States of America* **118.** issn: 10916490. (2022) (Dec. 2021).

11. Baran, Y. *et al.* MetaCell: Analysis of Single-Cell RNA-seq Data Using K-nn Graph Partitions. *Genome Biology* **20,** 206. issn: 1474-760X. (2023) (Oct. 2019).

12. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential Abundance Testing on Single-Cell Data Using k-Nearest Neighbor Graphs. *Nature Biotechnology* **40,** 245–253. issn: 1087-0156, 1546-1696. (2023) (Feb. 2022).

13. La Manno, G. *et al.* RNA Velocity of Single Cells. *Nature* **560,** 494–498. issn: 1476-4687. (2024) (Aug. 2018).

14. Salmen, F. *et al.* High-Throughput Total RNA Sequencing in Single Cells Using VASA-seq. *Nature Biotechnology* **40,** 1780–1793. issn: 1546-1696. (2024) (Dec. 2022).

15. Stoeckius, M. *et al.* Simultaneous Epitope and Transcriptome Measurement in Single Cells. *Nature Methods* **14,** 865–868. issn: 1548-7105. (2024) (Sept. 2017).

16. Schäbitz, A. *et al.* Spatial Transcriptomics Landscape of Lesions from Non-Communicable Inflammatory Skin Diseases. *Nature Communications* **13,** 7729. issn: 2041-1723. (2024) (Dec. 2022).