Immune infiltrate estimation: review of methods for deriving cell type profiles from purified cell sample data

Maxim Zaslavsky

June 2016

1 Introduction

Several groups have introduced methods for the deconvolution of bulk tumor gene expression data. Every method includes a unique procedure for isolating a reference profile corresponding to each cell type using sample gene expression data collected from enriched cell lines. For robust deconvolution, it is essential that these reference profiles – taking the form of marker gene lists or characteristic gene expression vectors – are determined carefully. Specifically, we are interested in the following properties of these methods:

- Do these training methods extract biological intuition or noise?
- How well do their selected genes or expression vectors differentiate between similar classes?
- How well can these methods differentiate all classes, based on metrics like condition number?
- Do their chosen genes overlap, and are the differences between their training expression vectors or marker genes biologically significant or noise?
- How unique are genes to individual cell types? How many are shared between multiple types?

We introduce each approach, discuss its theoretical limitations, and examine its output to understand whether there is motivation for new approaches.

2 Methods for selecting reference profiles

2.1 Marker gene methods

There are three notable methods that produce marker gene lists. First, [1] examined gene expression within immune cell types and in other tissue to produce a list of genes that are specifically expressed in particular immune cell types. To determine whether a certain gene g is exemplary of any cell type(s), the authors find the array with the highest expression level of g and determine which enriched cell type it corresponds to. They multiply this highest expression level by 0.1625 (chosen arbitrarily) and add the maximum expression level of g seen in non-immune tissue samples. If this weighted sum is greater than the next highest expression level of gene g across arrays from other immune cell types, then g is considered characteristic of the cell type in which it was most highly expressed. Finally, if this gene has higher expression in another immune cell type than this weighted sum, the gene is also considered to be characteristic of the other cell type.

However, fold change alone is a poor indicator of uniqueness; high expression should not be the only indicator that a gene corresponds to a particular cell type! A more robust method would consider rare expression, even at low levels.

[2] pursues the same task with a similar method. To determine whether the expression of gene g is

characteristic of cell type t, the authors essentially compute the score:

$$\Delta_{g,t} = \min_{e_i \in X_t} (e_i(g)) - \max_{t' \in T - \{t\}} (\operatorname{mean}_{e_j \in X_{t'}}(e_j(g))),$$

where X_i is the set of all arrays of cell type i and e(g) is the expression of g in some array. Then, they keep all g's with highest $\Delta_{q,t}$. The authors do not specify their filtering cutoff, unfortunately. The authors finally add some cell type-specific genes for populations not sampled, again without much detail (the code is not available).

This filtering mechanism ensures that selected genes are unique to their corresponding cell types, but would fail if any two types are very similar. In this case, genes whose differential expression has biological meaning but is small might not pass the filter, whereas genes with seemingly high differential expression – as may be found in the noise from low sample sizes – may pass.

Finally, though [3] attempts to estimate tumor purity, which is the absolute fraction of stromal and immune cells in a tumor sample, instead of the relative abundances of specific immune infiltrate cell types, the method also relies on identifying immune signature genes from gene expression in enriched samples and thus deserves investigation. The authors simply divided samples into extremely low and extremely high immune cell infiltration groups (using leukocyte methylation signature scores that are given in many TCGA datasets), removing any samples with medium immune cell infiltration. They computed Significance Analysis of Microarray (SAM) scores on the differential expression of genes between the high-and low-infiltration groups [4]. They selected genes that were significantly differentially expressed to form a gene list.

SAM is a straightforward, well-known, and statistically sound method for finding genes that are differentially expressed between two classes. Moreover, the SAM technique can be applied to multi-class situations to determine genes that are significantly differentially expressed in one combination of cell types versus another. I believe SAM would form more robust gene lists in comparison to previous methods that are based solely on fold change.

2.1.1 Analysis

Our first measure of whether these marker gene extraction methods are successful is whether known immune pathways are enriched in the gene list. For example, are the genes that these methods believe to be associated with T cells part of the T cell receptor signaling pathway, or are these methods pulling out noise?

The two marker gene lists, which we call IRIS [1] and Bindea [2], do not have much agreement on B cells or T cells; on NK cells, there are no intersecting genes at all. We run gene ontology enrichment analysis on the genes they have in common and on the genes unique to each list to see which T cell pathways are found and where. The resulting significant (p < .001) GO terms are contained in the tables below. Though the genes are different, they belong to the same pathways. The IRIS list contains much more noise than the Bindea list. This suggests that the method of ??? is more effective at extracting the unique properties of each immune cell subtype.

GO terms in intersection of T cell IRIS and Bindea marker gene lists:

1.	T cell receptor signaling pathway	7. T cell aggregation
2.	antigen receptor-mediated signaling pathway	8. lymphocyte aggregation

3. T cell costimulation

6. T cell activation

- 4. lymphocyte costimulation
- 5. immune response-activating cell surface receptor signaling pathway
- lymphocyte aggrega
- 9. leukocyte aggregation
- 10. T cell selection
- 11. leukocyte cell-cell adhesion
- 12. immune response-activating signal transduction
- 13. positive regulation of T cell activation

- 14. homotypic cell-cell adhesion
- 15. positive regulation of homotypic cell-cell adhesion
- 16. positive regulation of leukocyte cell-cell adhesion
- 17. immune response-regulating cell surface receptor signaling pathway
- 18. activation of immune response
- 19. positive regulation of cell-cell adhesion
- 20. lymphocyte activation
- 21. positive regulation of lymphocyte activation
- 22. positive regulation of leukocyte activation
- 23. immune response-regulating signaling pathway
- 24. positive regulation of immune response
- 25. T cell differentiation in thymus
- 26. thymocyte aggregation
- 27. positive regulation of cell activation
- 28. regulation of T cell activation
- 29. regulation of leukocyte cell-cell adhesion
- 30. leukocyte activation
- 31. regulation of homotypic cell-cell adhesion
- 32. single organismal cell-cell adhesion
- 33. single organism cell adhesion
- 34. cell adhesion

- 35. biological adhesion
- 36. positive regulation of cell adhesion
- 37. regulation of lymphocyte activation
- 38. regulation of cell-cell adhesion
- 39. cell-cell adhesion
- 40. positive regulation of immune system process
- 41. regulation of leukocyte activation
- 42. cell activation
- 43. regulation of cell activation
- 44. regulation of immune response
- 45. thymic T cell selection
- 46. positive T cell selection
- 47. positive regulation of calcium-mediated signaling
- 48. T cell differentiation
- 49. regulation of cell adhesion
- 50. regulation of calcium-mediated signaling
- 51. regulation of immune system process
- 52. immune system process
- 53. lymphocyte differentiation
- 54. immune response
- 55. olfactory bulb axon guidance
- 56. positive regulation of response to stimulus

GO terms of T cell genes in IRIS but not in Bindea:

- 1. cell division
- 2. nuclear division
- 3. organelle fission
- 4. cell cycle process
- 5. mitotic nuclear division
- 6. mitotic cell cycle process
- 7. mitotic cell cycle
- 8. cell cycle
- 9. cell cycle checkpoint
- 10. mitotic cell cycle checkpoint
- 11. G2/M transition of mitotic cell cycle

- 12. cell cycle G2/M phase transition
- 13. anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
- 14. T cell activation
- 15. T cell aggregation
- 16. lymphocyte aggregation
- 17. leukocyte aggregation
- 18. regulation of cell cycle
- 19. spindle organization
- 20. mitotic cell cycle phase transition
- 21. leukocyte cell-cell adhesion

- 22. regulation of mitotic cell cycle
- 23. cell cycle phase transition
- 24. negative regulation of mitotic cell cycle
- 25. regulation of spindle organization
- 26. homotypic cell-cell adhesion
- 27. mitotic spindle organization
- 28. somatic diversification of T cell receptor genes
- 29. somatic recombination of T cell receptor gene segments
- 30. T cell receptor V(D)J recombination
- 31. spindle stabilization
- 32. spindle assembly involved in meiosis
- 33. lymphocyte activation
- positive regulation of ubiquitin-protein transferase activity
- 35. regulation of ubiquitin homeostasis
- 36. free ubiquitin chain polymerization
- 37. positive regulation of ligase activity
- 38. meiotic cell cycle
- 39. mitotic nuclear envelope disassembly
- 40. membrane disassembly
- 41. nuclear envelope disassembly
- 42. forebrain neuroblast division
- 43. leukocyte activation
- 44. sister chromatid segregation
- 45. response to insecticide
- 46. activation of anaphase-promoting complex activity
- 47. single organismal cell-cell adhesion
- 48. neural precursor cell proliferation
- 49. regulation of cell cycle process
- 50. cell proliferation
- 51. meiotic spindle organization
- 52. cell activation
- 53. regulation of ligase activity
- 54. regulation of ubiquitin-protein transferase activity
- 55. histone-serine phosphorylation

- 56. neuronal stem cell division
- 57. neuroblast division
- 58. single organism cell adhesion
- 59. microtubule cytoskeleton organization
- 60. V(D)J recombination
- 61. immune system development
- 62. meiotic nuclear division
- 63. mitotic G2 DNA damage checkpoint
- 64. interleukin-5 production
- 65. regulation of interleukin-5 production
- 66. meiotic cell cycle process
- 67. DNA integrity checkpoint
- 68. negative regulation of mitotic cell cycle phase transition
- 69. nuclear envelope organization
- 70. positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition
- positive regulation of proteolysis involved in cellular protein catabolic process
- 72. regeneration
- 73. oogenesis
- 74. spindle assembly
- 75. organ regeneration
- 76. T cell costimulation
- 77. positive regulation of protein ubiquitination
- 78. nuclear chromosome segregation
- 79. lymphocyte costimulation
- 80. cell-cell adhesion
- 81. centrosome localization
- 82. regulation of mitotic spindle organization
- 83. negative regulation of cell cycle phase transition
- 84. positive regulation of cellular protein catabolic process
- 85. negative regulation of cell cycle
- 86. positive regulation of protein modification by small protein conjugation or removal
- 87. negative regulation of cell division
- 88. single-organism organelle organization

GO terms of T cell genes in Bindea but not in IRIS:

- 1. T cell receptor signaling pathway
- 2. antigen receptor-mediated signaling pathway
- 3. positive regulation of immune system process
- 4. regulation of immune system process
- 5. positive regulation of leukocyte activation
- 6. regulation of immune response
- 7. positive regulation of cell activation
- 8. regulation of T cell activation
- 9. regulation of leukocyte cell-cell adhesion
- 10. regulation of homotypic cell-cell adhesion
- 11. regulation of cell adhesion
- immune response-activating cell surface receptor signaling pathway
- 13. positive regulation of immune response
- 14. positive regulation of cell adhesion
- 15. regulation of lymphocyte activation
- 16. regulation of cell-cell adhesion
- 17. regulation of leukocyte activation
- 18. T cell activation
- 19. T cell aggregation
- 20. lymphocyte aggregation
- 21. leukocyte aggregation
- 22. cell adhesion
- 23. biological adhesion
- 24. regulation of cell activation
- 25. positive regulation of T cell activation
- 26. positive regulation of homotypic cell-cell adhesion
- 27. positive regulation of leukocyte cell-cell adhesion
- 28. leukocyte cell-cell adhesion
- 29. immune response-activating signal transduction
- 30. homotypic cell-cell adhesion
- 31. immune response-regulating cell surface receptor signaling pathway
- 32. activation of immune response

- 33. positive regulation of cell-cell adhesion
- 34. immune response
- 35. positive regulation of lymphocyte activation
- 36. T cell costimulation
- 37. lymphocyte costimulation
- 38. immune system process
- 39. lymphocyte activation
- 40. immune response-regulating signaling pathway
- 41. positive regulation of interleukin-2 biosynthetic process
- 42. leukocyte activation
- 43. single organismal cell-cell adhesion
- 44. single organism cell adhesion
- 45. regulation of interleukin-2 biosynthetic process
- 46. interleukin-2 biosynthetic process
- 47. cell-cell adhesion
- 48. cell activation
- 49. regulation of defense response to virus by virus
- 50. positive regulation of interleukin-2 production
- 51. T cell differentiation
- 52. positive regulation of response to stimulus
- 53. regulation of interleukin-2 production
- 54. positive regulation of alpha-beta T cell activation
- 55. interleukin-2 production
- 56. positive regulation of cytokine biosynthetic process
- 57. positive regulation of myeloid dendritic cell activation
- 58. lymphocyte differentiation
- 59. positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
- 60. regulation of alpha-beta T cell activation
- 61. positive regulation of lymphocyte mediated immunity

- 62. Fc-epsilon receptor signaling pathway
- 63. positive regulation of signal transduction
- 64. positive regulation of adaptive immune response

2.2 Expression barcodes

Storing representative gene expression signatures, as opposed to just marker genes, is key to more robust predictions of immune infiltrate cell type abundances. These distinctive transcriptional profiles are often called unique expression "barcodes" (seemingly named for the heatmaps commonly used to visualize microarray data). We now examine two methods that extract representative expression profiles.

[5] introduces the following procedure to select barcodes. For each expressed gene, the authors find the two cell types with highest expression of this gene (perhaps in terms of mean expression across all samples from each cell type, although the details are not given). If the gene is differentially expressed within a 95% fold change confidence interval between those cell types, the gene is flagged as a potential marker for the cell type with higher expression. This approach would clearly fail for very similar subtypes, and may only pull out noise because of low sample sizes. So the authors also compare the cell types with highest and third-highest expression of this gene in case it is hard to tell between the top two groups. They progressively refine their basis matrix with an increasing number of top genes, and report that they minimize the condition number of their matrix with an intermediate number of included genes (360 genes).

The authors note that their method produces a well-conditioned matrix. This is an important consideration because the condition number, defined as the ratio of the largest to smallest singular values in the singular value decomposition of the basis matrix, estimates how imprecise solutions to linear systems with this matrix are, and thus is a good proxy for the accuracy of deconvolution under the well-justified biological assumption of linearity [6]. The smaller the condition number, the better conditioned the basis matrix is, meaning the cell types are more distinct. However, more strict statistical testing with a controlled false discovery rate is desired.

[7] provides this desired statistical rigor. Like the previous method, this one also iteratively deletes irrelevant genes. The authors find significantly differentially expressed genes between all populations using two-sided unequal variance *t*-tests, with a (fairly loose) false discovery rate threshold of q < .3 and with log fold change greater than 2.0. The number of selected genes per cell type is reduced from at most the first 150 towards 50 final selected genes in search of the best-conditioned matrix (minimum condition number).

Here is an example of the output of these methods. [5] provides raw samples from several populations: T cells, two lines of B cells, and monocytes. Figure 1 and 2 are correlation matrices of the pure samples and of processed basis matrices (via [7] codebase), respectively. Note the poor differentiation in the raw data (especially note the scale), whereas differentiation is much easier in the processed matrix.

2.2.1 Analysis

We want to characterize how well expression barcode methods distinguish similar cell types. [5] does not provide code to regenerate their full basis matrix from many samples. However, I was able to reproduce the basis matrix from [7] using their tools and supplied input data, albeit with less filtering: the authors postprocessed their signature matrix to remove some junk genes using annotations from cancer cell lines. Though my basis matrix thus included more genes, I obtained a very similar condition number to their matrix (which they call LM22), and the genes in common all had almost exactly the same expressions throughout. This suggest that the postprocessing that was poorly described and that I was unable to run did not significantly refine the matrix.

I performed hierarchical clustering and computed pairwise Pearson correlations between cell typespecific profiles in the signature matrices from [5] and [7]. The pairwise Pearson correlation of the LM22 matrix [7] showed nice differentiation between cell types, and biologically-related cell types were highly correlated (Figure 3). In contrast, the pairwise Pearson correlations from the matrix in [5], hereafter called Abbas, showed very poor differentiation among several B cell types (Figure 4). I also computed pairwise Pearson



Figure 1: Pairwise correlation in raw data from [5].



Figure 2: Pairwise correlation in basis matrix created from raw data of [5].

correlations from the combined matrices (Figure 5). Different methods with different datasets still produce nice expected correlations, although are also several unexpected inter-matrix correlations.



Figure 3: Pairwise Pearson correlation in LM22 [7].

Hierarchical clustering of genes and cell types in LM22 generally recovers biological similarities between cell types (Figure 6). There is one exception: gamma delta T cells. However, this cell type has been flagged as problematic and may be ignored [8].

Since LM22 has nice differentiation between cell types, it is interesting to examine the most similar cell types in this matrix. The Pearson correlations and the hierarchical clustering reveal that the following classes in LM22 are most similar:

- B cells memory, naive
- CD4 T cells naive, memory resting

When one of each pair of similar cell types is removed, the condition number decreases from 11.38 to 9.30, meaning the resulting matrix is considerably better at deconvolving the more distinct set of cell types.

3 Future directions

In total, these papers have 390 microarrays samples. I downloaded and normalized all this array data. We can construct a much richer set of expression profiles from this expanded dataset. In fact, the sample size could potentially allow us to model variance and not just use mean expression profiles, which could be critical for deconvolving the immune contexture of tumors, in which immune cells may have differing activations or other properties depending on the state of the tumor.



Figure 4: Pairwise correlation in Abbas [5].



Figure 5: Pairwise pearson correlation in combination of LM22 and Abbas basis matrices, as well as with raw data from [5].



Figure 6: LM22 hierarchical clustering

Since RNAseq is popular today for tumor sequencing, it is desirable to obtain enriched immune cell line RNAseq data and produce a new basis matrix. However, online discussion suggests that RNAseq does not support the independence assumptions in microarray analysis: https://www.biostars.org/p/160961/. This context may require different reference profile expression methods.

References

1. Abbas A, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (iRIS): Immune-specific genes identified from a compendium of microarray expression data. Genes and immunity. Nature Publishing Group; 2005;6: 319–331.

2. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. Elsevier; 2013;39: 782–795.

3. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. Nature Publishing Group; 2013;4.

4. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2005;102: 12837–12842.

5. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PloS one. Public Library of Science; 2009;4: e6098.

6. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type–specific gene expression differences in complex tissues. Nature methods. Nature Publishing Group; 2010;7: 287–289.

7. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. Nature Publishing Group; 2015;12: 453–457.

8. Senbabaoglu Y, Winer AG, Gejman RS, Liu M, Luna A, Ostrovnaya I, et al. The landscape of t cell infiltration in human cancer and its association with antigen presenting gene expression. bioRxiv. Cold Spring Harbor Labs Journals; 2015; doi:10.1101/025908