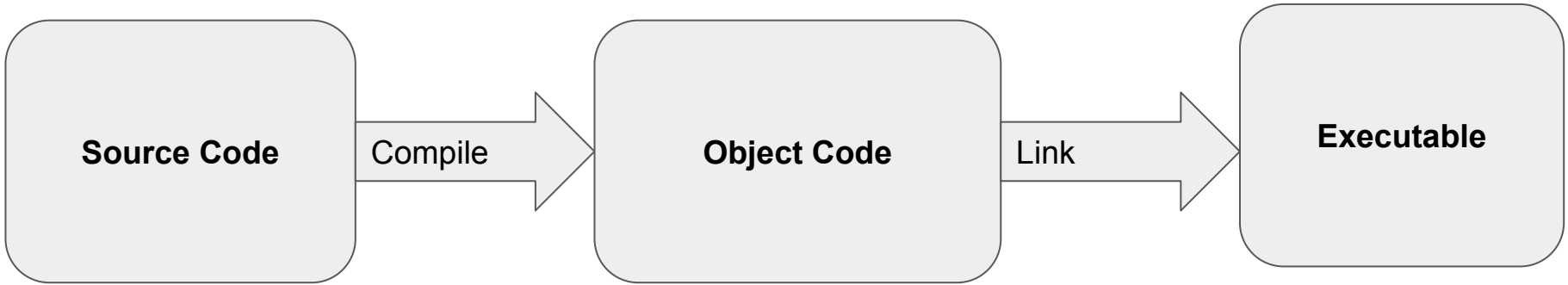
ATCGes - A Tool for Curating Genome Expression Signatures

By: Louie Dinh

Biology!

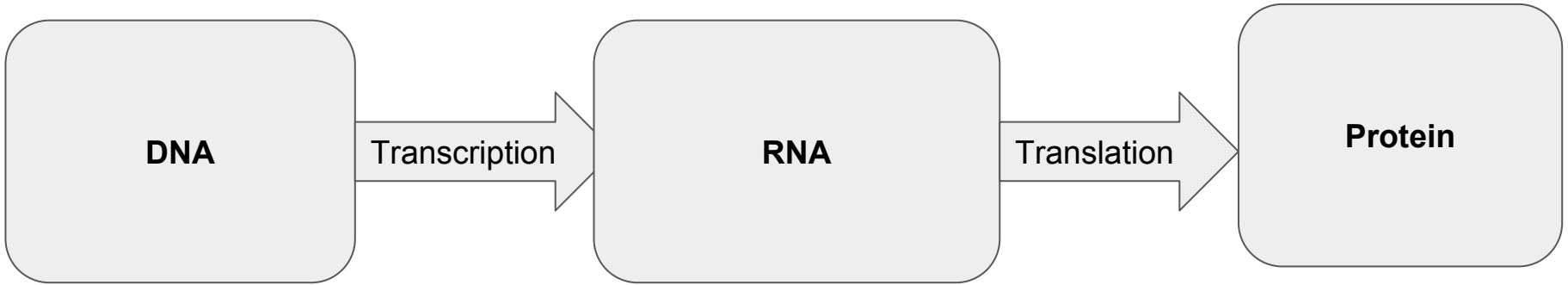
Central Dogma [For Computer Scientists]

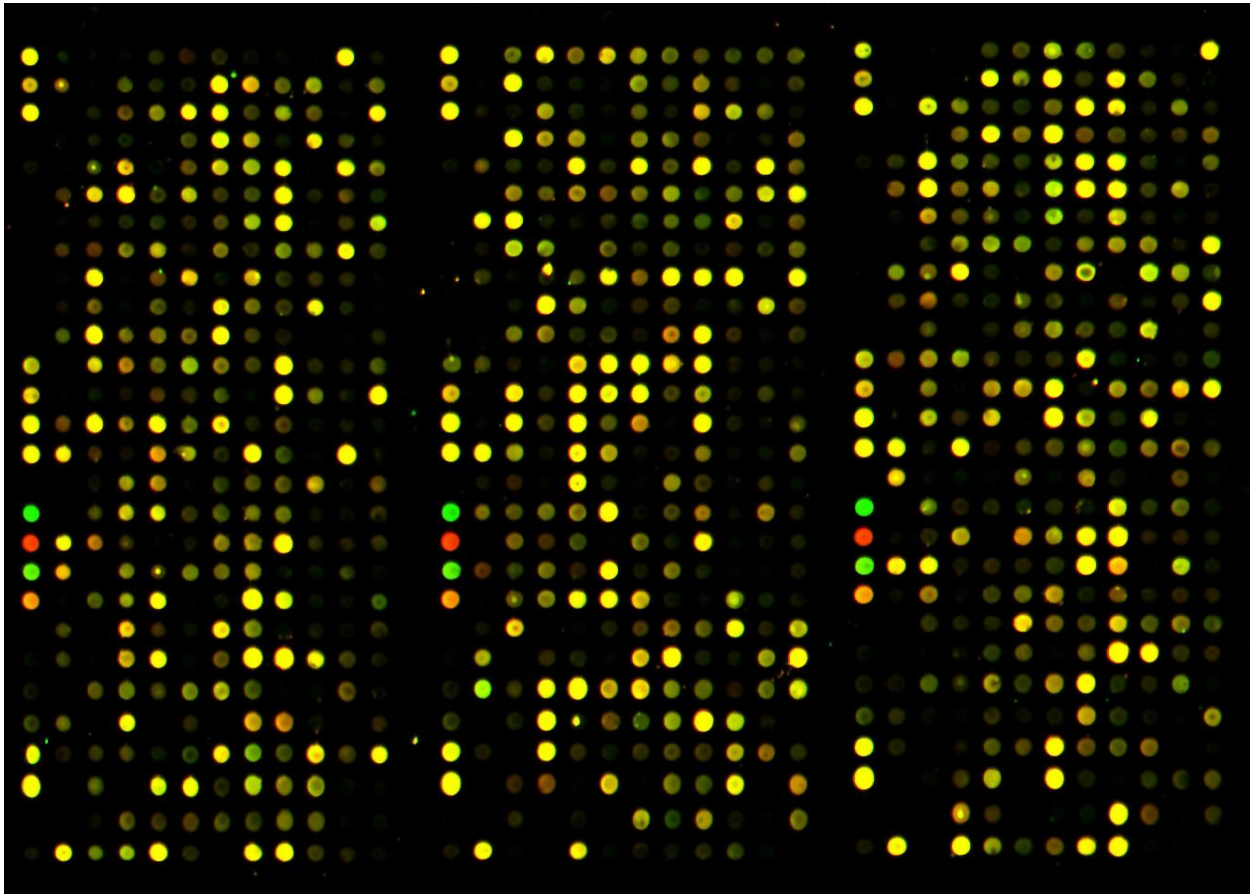
- REVIEW: Compiling Source Code



Central Dogma [For Biologists]

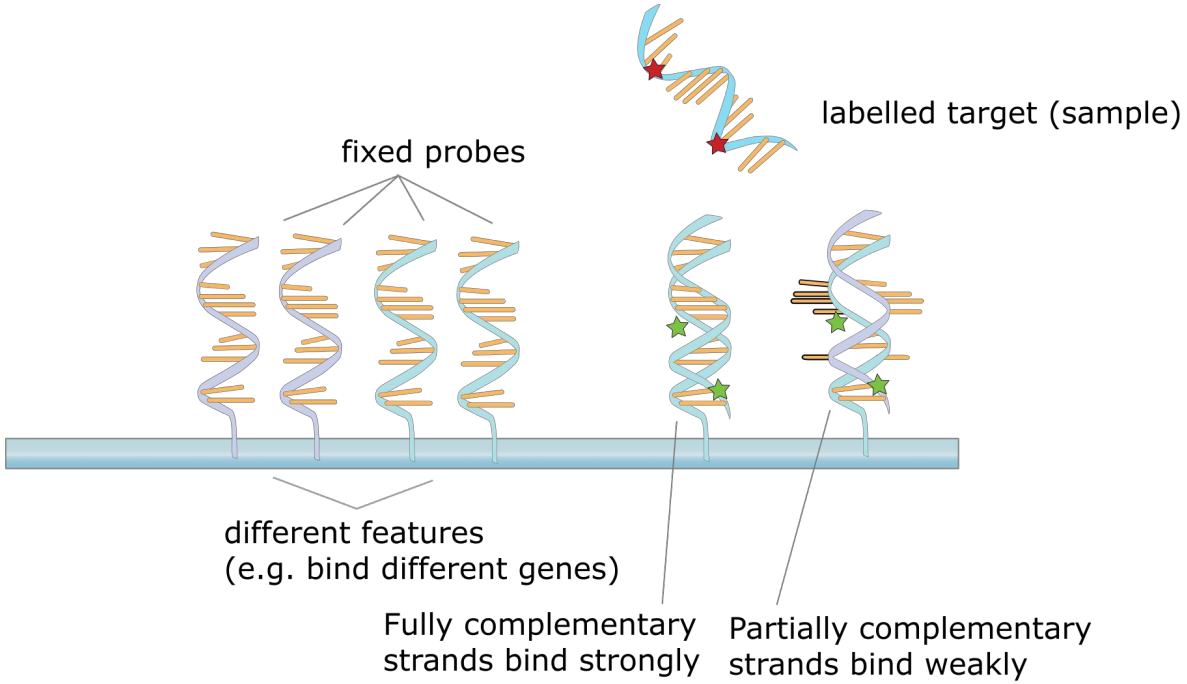
- NEW: Compiling Life





Microarrays - Measure all the genes!

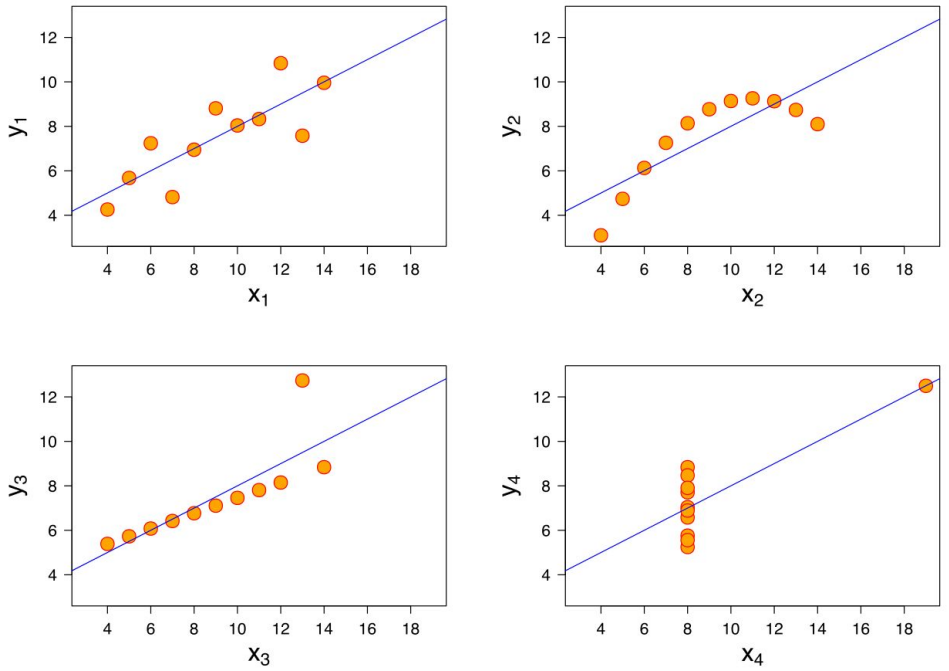
~~Magnets~~ Microarrays - How Do They Work?!



Problem!

Microarray Data Measures Raw Abundance

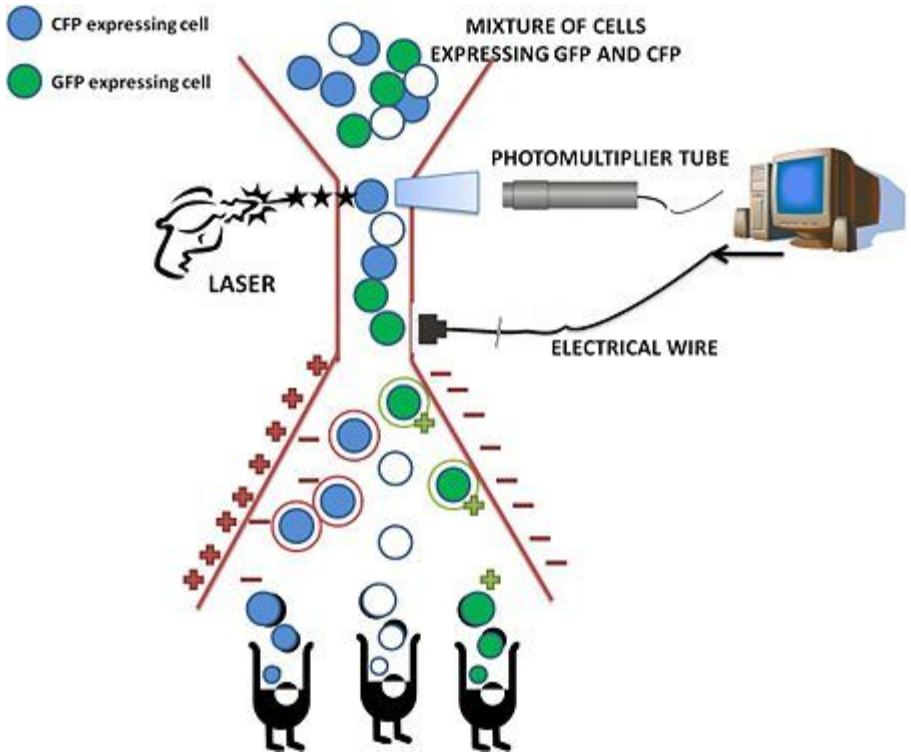
- When comparing two samples, is difference from No. of cells or strength of expression?
- How do we account for genes that are just used more? i.e higher background signal?
- How do you figure out the per-cell-type contribution?
- Confounding Factors



Solution!

Deconvolution

- Mixed sample is being measured
- Figure out how to unmix them
- Can either be done physically (FACS, flow cytometry) or computationally
- Computational deconvolution **requires a signature matrix**, a set of genes that uniquely identifies each cell



Identify a set of biomarkers (genes) that can distinguish between the samples (cells).

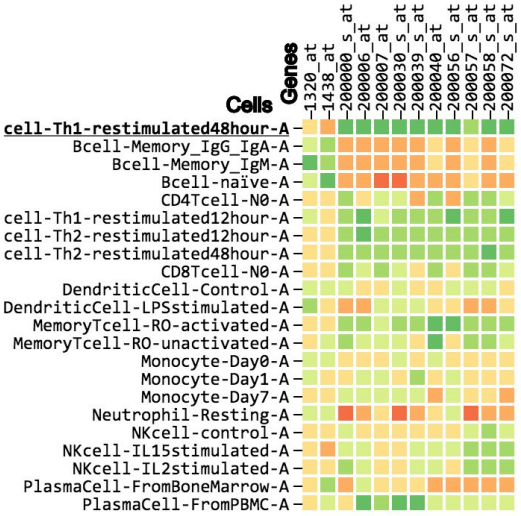
Data Abstraction

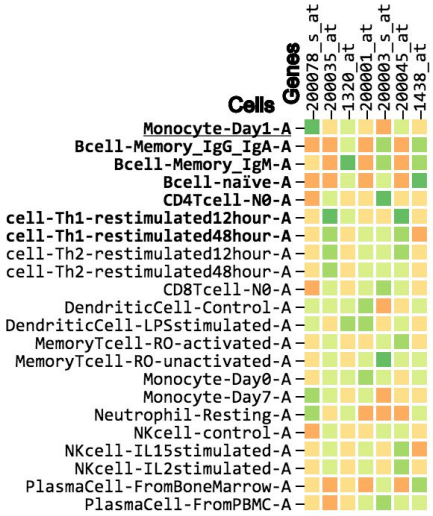
- Tabular data
  - Key = gene_id, cell_id, replicate_no. [Categorical]
  - Value = Luminance Measure [Quantitative]
- Scales
  - ~20k genes
  - ~2 dozen cells
  - ~ 5 replicates
  - Luminance between 0-100,000

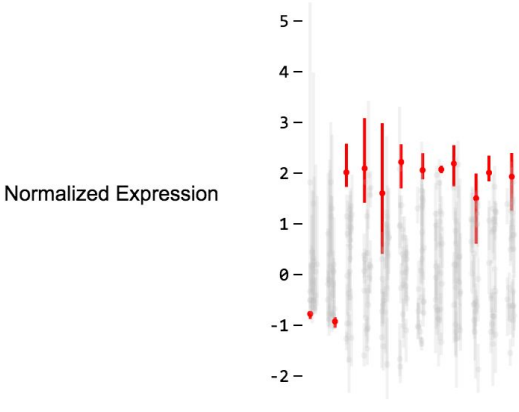
Task

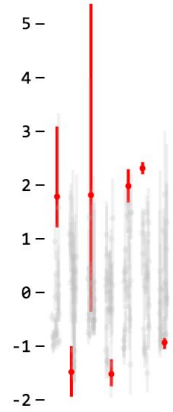
- For each cell, identify a gene that is differentially expressed when compared to other cells.
- The ideal biomarker has low variance, and whose signal doesn’t overlap with the other cells.

ATCGes - A Tool For Curating Gene Expression Signatures

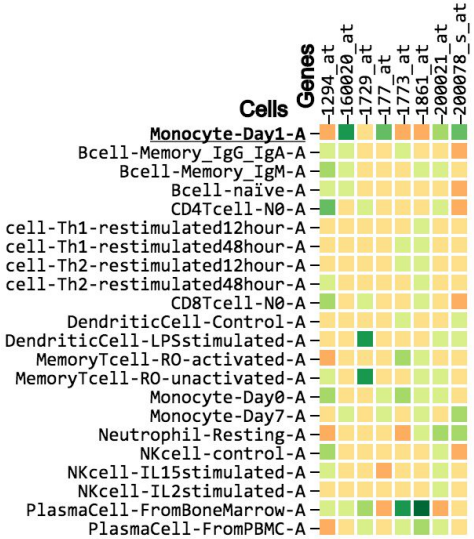


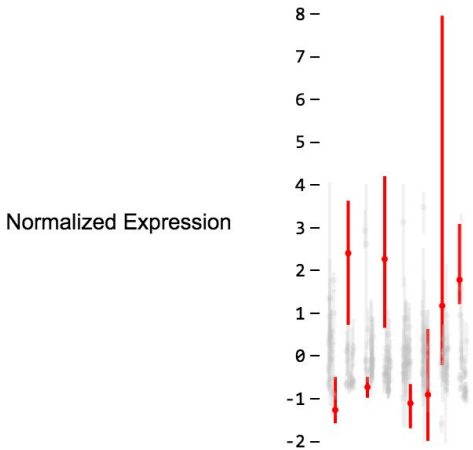


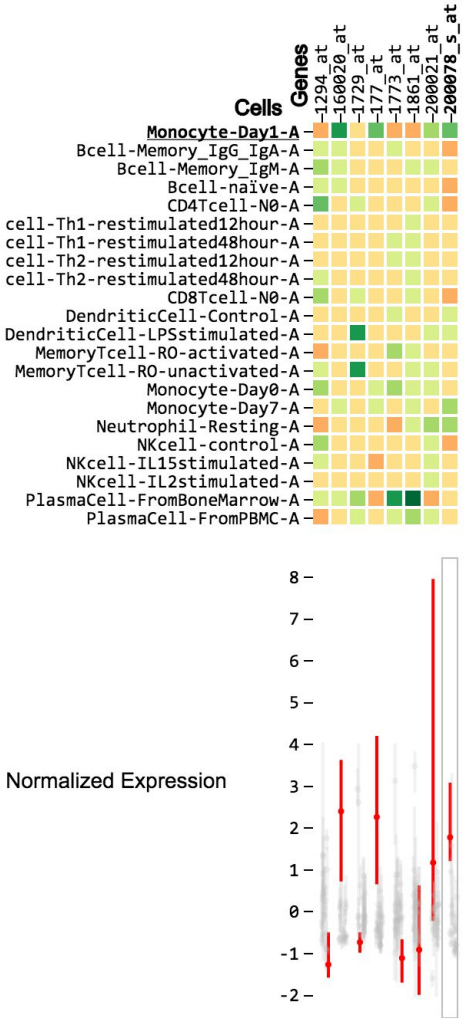


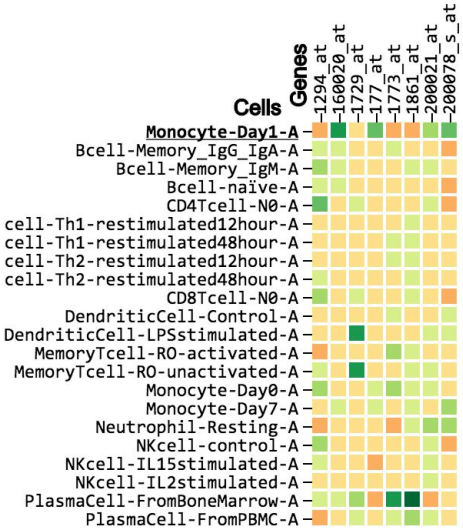


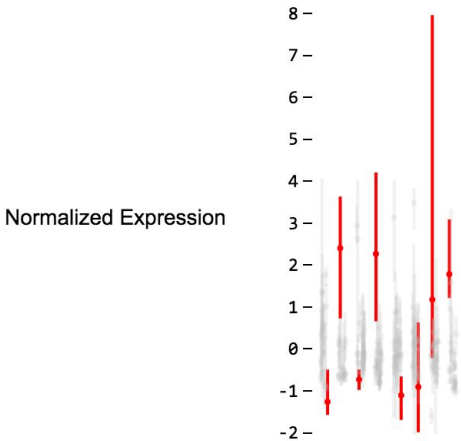
Demo!

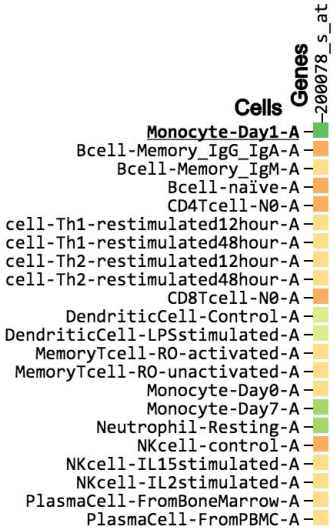

Genes
Cells
1294_at
160020_at
1729_at
177_at
1773_at
1861_at
200021_at
200078_s_at
Monocyte-Day1-A
Bcell-Memory_IgG_IgA-A
Bcell-Memory_IgM-A
Bcell-naïve-A
CD4Tcell-N0-A
cell-Th1-restimulated12hour-A
cell-Th1-restimulated48hour-A
cell-Th2-restimulated12hour-A
cell-Th2-restimulated48hour-A
CD8Tcell-N0-A
DendriticCell-Control-A
DendriticCell-LPSstimulated-A
MemoryTcell-RO-activated-A
MemoryTcell-RO-unactivated-A
Monocyte-Day0-A
Monocyte-Day7-A
Neutrophil-Resting-A
NKcell-control-A
NKcell-IL15stimulated-A
NKcell-IL2stimulated-A
PlasmaCell-FromBoneMarrow-A
PlasmaCell-FromPBMC-A


Normalized Expression
8
7
6
5
4
3
2
1
0
-1
-2


Cells
Genes
1294_at
16020_at
1729_at
177_at
1773_at
1861_at
200021_at
200078_s_at
Monocyte-Day1-A
Bcell-Memory_IgG_IgA-A
Bcell-Memory_IgM-A
Bcell-naïve-A
CD4Tcell-N0-A
cell-Th1-restimulated12hour-A
cell-Th1-restimulated48hour-A
cell-Th2-restimulated12hour-A
cell-Th2-restimulated48hour-A
CD8Tcell-N0-A
DendriticCell-Control-A
DendriticCell-LPSstimulated-A
MemoryTcell-RO-activated-A
MemoryTcell-RO-unactivated-A
Monocyte-Day0-A
Monocyte-Day7-A
Neutrophil-Resting-A
NKcell-control-A
NKcell-IL15stimulated-A
NKcell-IL2stimulated-A
PlasmaCell-FromBoneMarrow-A
PlasmaCell-FromPBMC-A
Normalized Expression
8
7
6
5
4
3
2
1
0
-1
-2

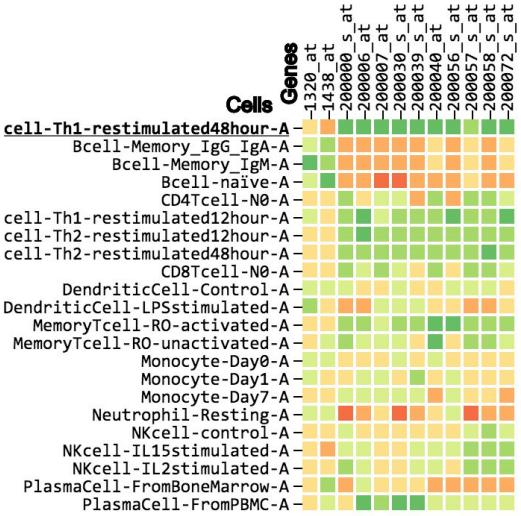

Cells
Genes
-1294_at
-160020_at
-1729_at
-177_at
-1773_at
-1861_at
-200021_at
-200078_s_at
Monocyte-Day1-A
Bcell-Memory_IgG_IgA-A
Bcell-Memory_IgM-A
Bcell-naïve-A
CD4Tcell-N0-A
cell-Th1-restimulated12hour-A
cell-Th1-restimulated48hour-A
cell-Th2-restimulated12hour-A
cell-Th2-restimulated48hour-A
CD8Tcell-N0-A
DendriticCell-Control-A
DendriticCell-LPSstimulated-A
MemoryTcell-RO-activated-A
MemoryTcell-RO-unactivated-A
Monocyte-Day0-A
Monocyte-Day7-A
Neutrophil-Resting-A
NKcell-control-A
NKcell-IL15stimulated-A
NKcell-IL2stimulated-A
PlasmaCell-FromBoneMarrow-A
PlasmaCell-FromPBMC-A

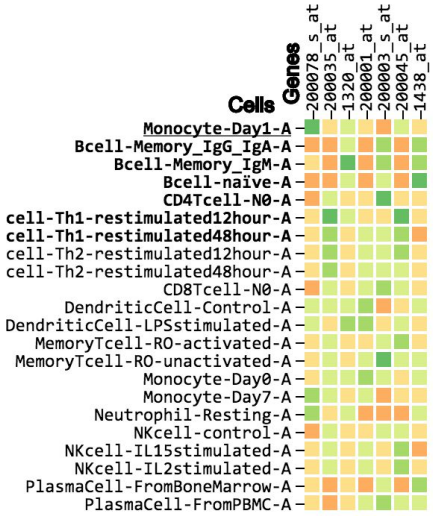

Normalized Expression
8
7
6
5
4
3
2
1
0
-1
-2

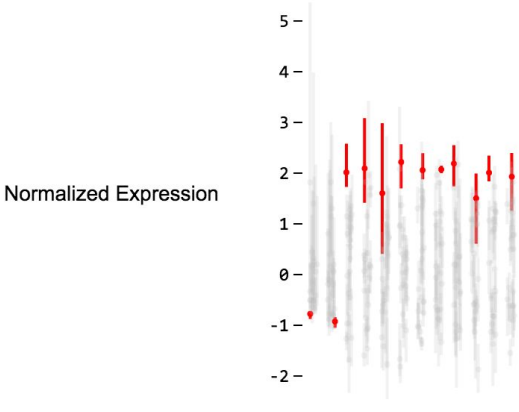

Cells
Genes
-200078_s_at
Monocyte-Day1-A
Bcell-Memory_IgG_IgA-A
Bcell-Memory_IgM-A
Bcell-naïve-A
CD4Tcell-N0-A
cell-Th1-restimulated12hour-A
cell-Th1-restimulated48hour-A
cell-Th2-restimulated12hour-A
cell-Th2-restimulated48hour-A
CD8Tcell-N0-A
DendriticCell-Control-A
DendriticCell-LPSstimulated-A
MemoryTcell-RO-activated-A
MemoryTcell-RO-unactivated-A
Monocyte-Day0-A
Monocyte-Day7-A
Neutrophil-Resting-A
NKcell-control-A
NKcell-IL15stimulated-A
NKcell-IL2stimulated-A
PlasmaCell-FromBoneMarrow-A
PlasmaCell-FromPBMC-A

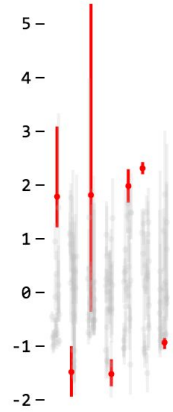

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5

ATCGes - A Tool For Curating Gene Expression Signatures









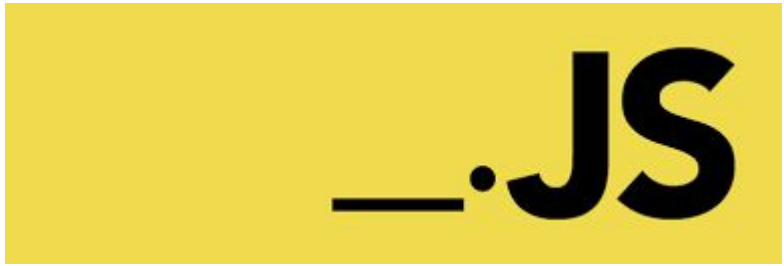
Implementation


R Studio


COLORBREWER 2.0
color advice for cartography


D3


GEO
Gene Expression Omnibus


_.JS

Strengths And Weaknesses

Strengths

- Filtering worked really well
- Visual occlusion was initially a big setback for the variance plot, but opacity makes it a great idiom.
- Partial solution is very powerful, allows for modifying the signature in an iterative fashion

Weaknesses

- Lots of Clicking
- Skewed expression signatures doesn’t use the full color range
- Expressions are scattered between the top and bottom of the range. Inversions would make it much more consistent in both the heatmap and the variance bars

Future Work

Future Work

- Integration into the full toolchain to benchmark against deconvolution output
- Better UI/UX with confirm/modify idiom rather than hunt and peck
- Drill-down + Roll-up capabilities to dig into a particular group of cells that are too similar. Re-normalizes the colormap so that we can actually separate out similar cells.

Lessons Learned and Conclusion

- Roadblocks are sometimes blessings in disguise
- Colour encoding is difficult but the correct scheme qualitatively changes your task effectiveness (at least feels that way).

- ACTGes met initial goals of curating a gene expression signature
- Focused on information density and fully presenting partial solution
- Iterative signature selection is a powerful paradigm

Thank You!