Behavior Approximation for Leukemia ceLL LineS (BALLLLS)

UCSD BISB Bootcamp 2013

Tony Aylward, Justin Huang, Yunjiang Qiu

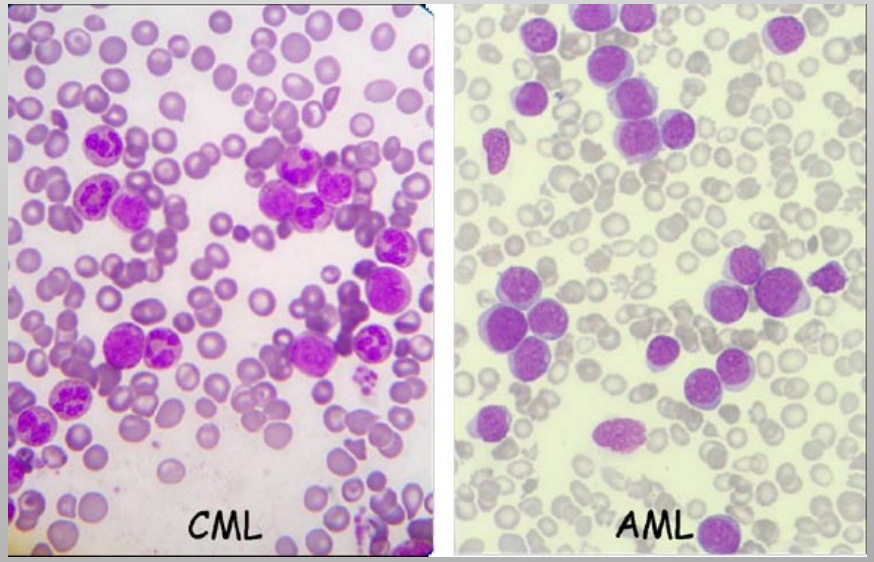
UC San Diego

Project Aims

1. Determine gene expression profiles for K562 (CML) and H1-ESC (embryonic stem cell) cell lines from RNA-seq data
2. Cluster gene expression profiles of patients generated from RNA-seq data by The Cancer Genome Atlas
3. Determine if K562 or H1-ESC cell lines can adequately represent AML in vitro

Background: Leukemia

- Chronic Myeloid Leukemia (CML)
  - Slow moving cancer
  - Cancer of differentiated blood cells
- Acute Myeloid Leukemia (AML)
  - Fast moving cancer
  - Abnormal cells accumulate inside bone marrow



Background: Cell Lines

- CML (K562)
  - From Source
  - RNA-seq data:
    - H1-ESC(GSM958737): Embryonic stem Cells
    - K562(GSM958731): 53 year old female patient
    - 80 AML female patients from TCGA (Expression calls already given)
- Processing

RNA-seq Alignment

- Alignment using Tophat

```
tophat --library-type fr-secondstrand -o /oasis/projects/nsf/csd399/serein/H1Rep1_test -p 16 /oas
is/projects/nsf/csd399/serein/bt2/hg19 /oasis/projects/nsf/csd399/serein/data/raw/wgEncodeCaltech
RnaSeqH1hescR1x75dFastqRep1.fastq
```

- Result

```
                Input:  25157723
               Mapped:  15247005 (60.6% of input)
          of these:    3717958 (24.4%) have multiple alignments (724 have >20)
60.6% overall read alignment rate.
```

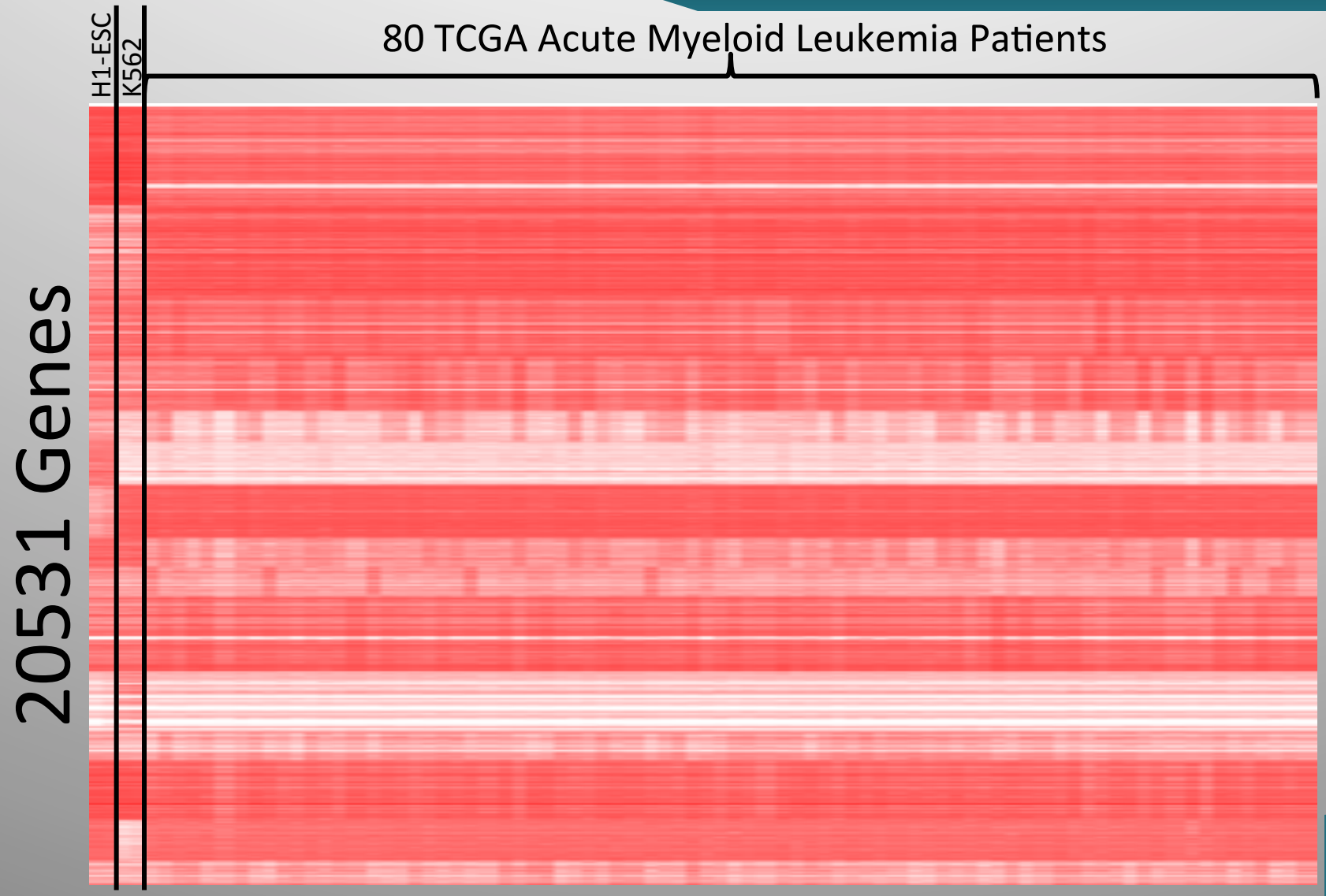
- RPKM
  - Reads per kilo base per million
  - (Read count * 1,000,000) / (total number of reads * kilo base of gene)
- Using DEGseq in R

```
Rscript DEGseq.R H1Rep2.bed hg19.refflat H1Rep2.exp
```

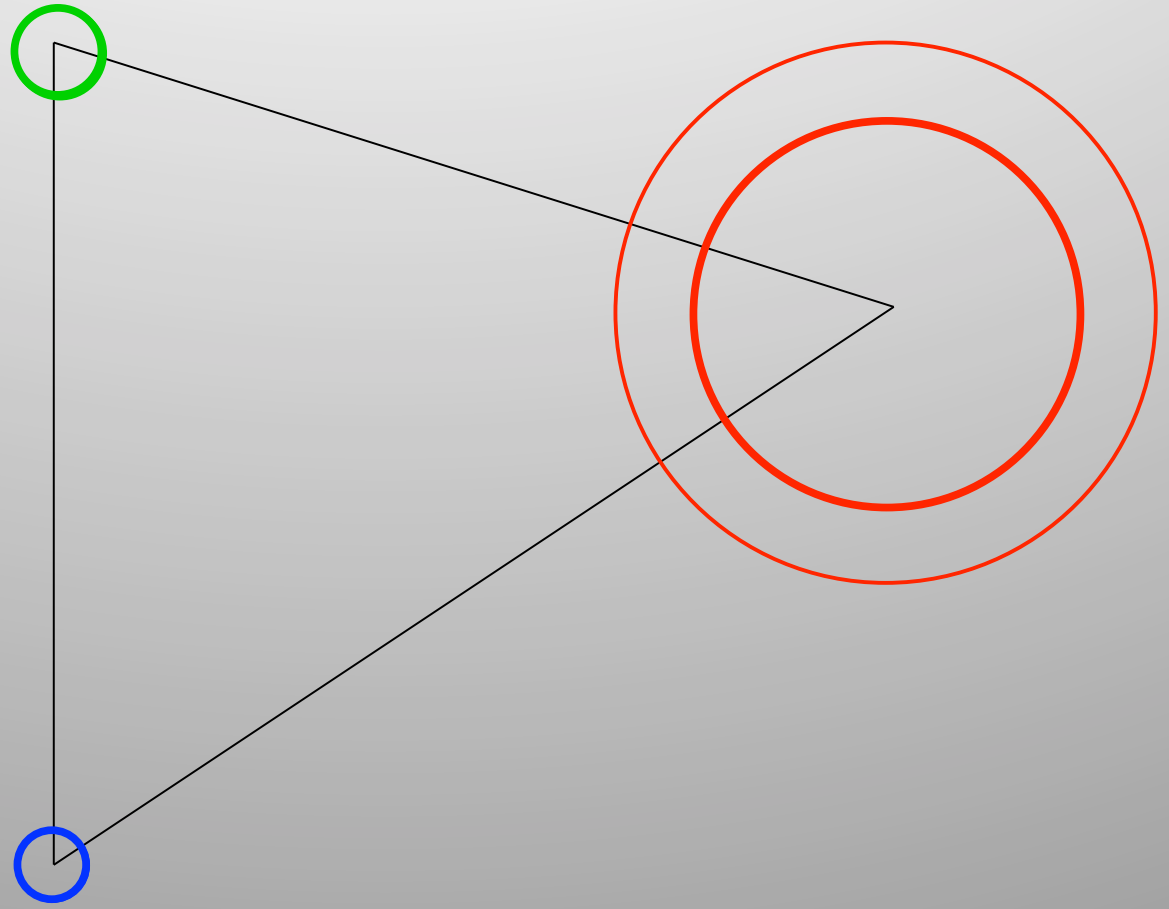
TCGA Patient Clustering

1. Data were normalized to a zero-one scale for each patient and cell line
2. A k-means clustering algorithm was used with k = 3. One profile from each cell line and one patient profile were chosen as the initial guesses for the means.
3. The following distance was used: d(x,y) = 1 – corr(x,y), where corr(x,y) is the Pearson correlation coefficient of profiles x and y.
4. Before clustering all data at once, we applied the algorithm to the patient profiles only. In that case, running the algorithm with k = 2 or k = 3 offered little qualitative improvement over running it with k = 1. We concluded that the patient profiles could safely be treated as a single cluster without subtypes.

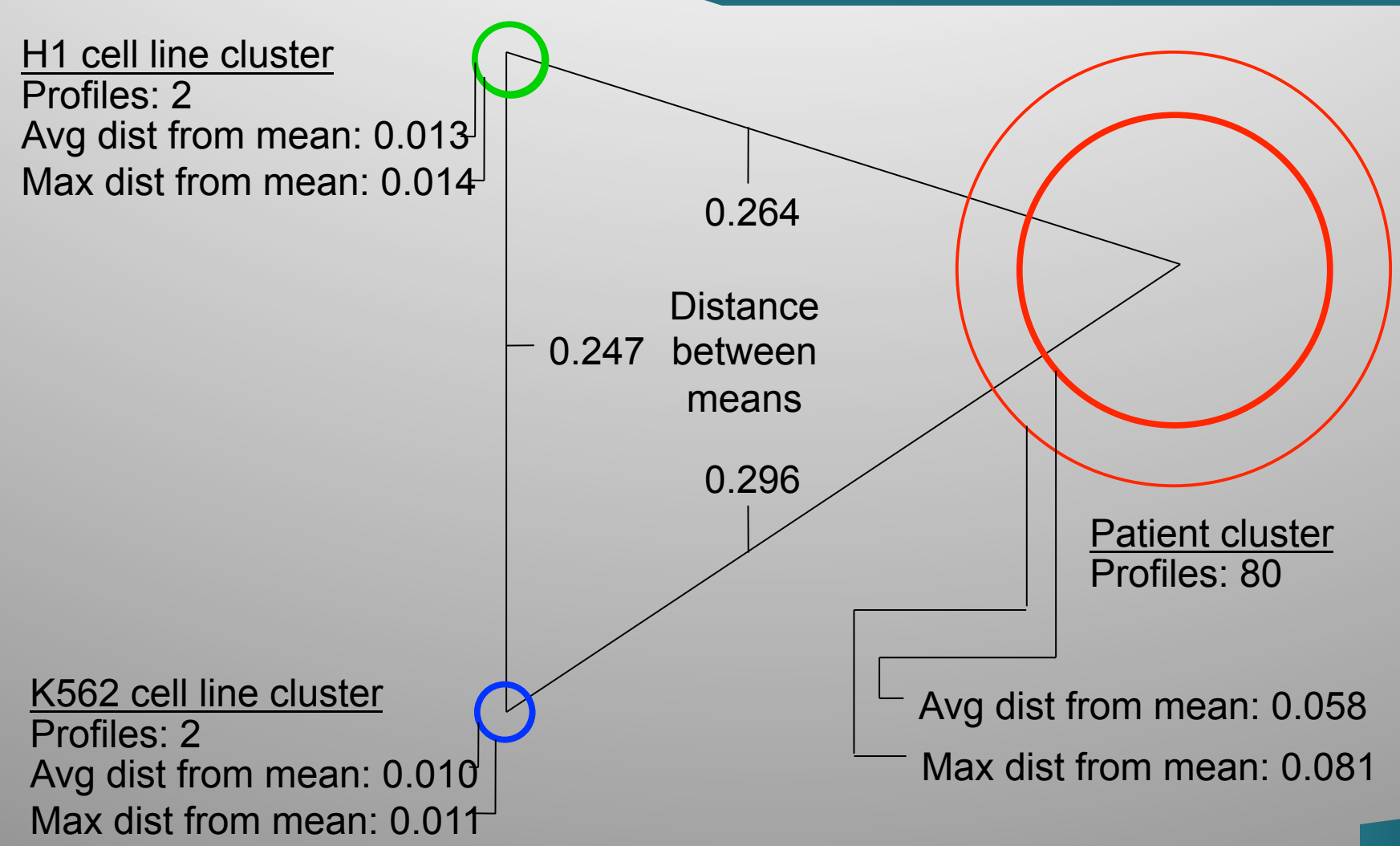
Patient Expression Profiles



Cell Line Similarity (Distances shown to scale)

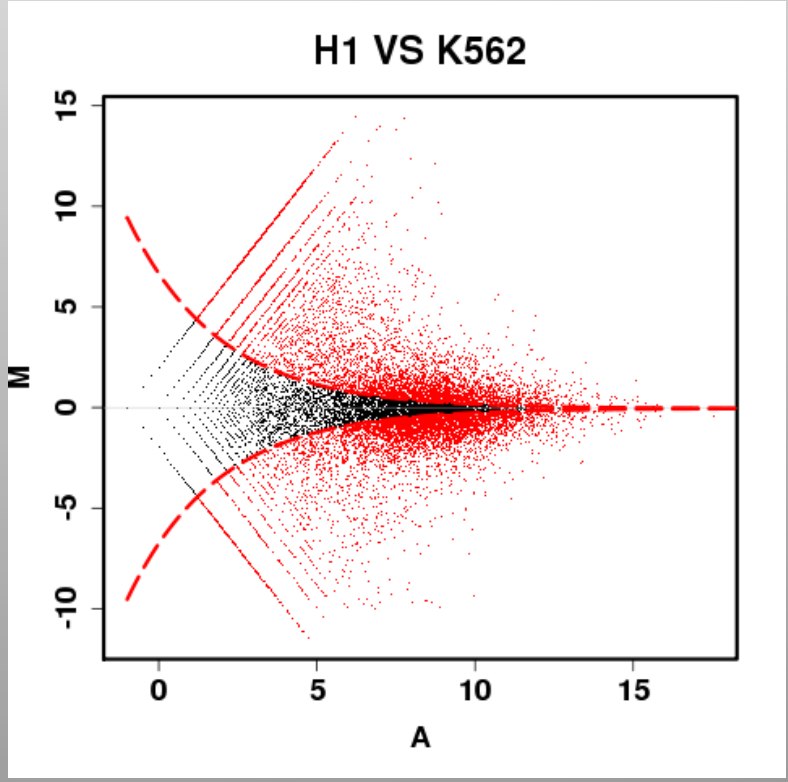


Cell Line Similarity (Distances shown to scale)



Differential expression analysis

- Search for genes that are differentially expressed between cell lines and patients
  - H1 vs K562
  - K562 vs Patients
  - H1 vs Patients
- Using DEGseq



Differentially Expressed Genes

	Patients vs K562	Patients vs H1-ESC	K562 vs H1-ESC
Number of Genes	9738 genes	10306 genes	7197 genes
Most Enriched Annotations	-Phosphoprotein -Alternative Splicing -Splice Variant -Zinc-Finger -Metal-Binding -Pleckstrin homology	-Alternative Splicing -Splice Variant -Phosphoprotein -Plasma Membrane Part -Golgi Apparatus -Mutagenesis Site	-Alternative Splicing -Splice Variant -Plasma Membrane Part -Biological Adhesion -Cell Adhesion -Phosphoprotein
Most Enriched KEGG Pathways	-Chemokine Signaling Pathway -Hematopoietic cell lineage -Lysosome -MAPK Signaling pathway	-Lysosome -Chemokine Signaling Pathway -MAPK Signaling pathway -B Cell Receptor Signaling Pathway	-Focal Adhesion -Cell Adhesion Molecules (CAMs) -Axon Guidance -Pathways in Cancer -ECM-Receptor Interaction

Conclusions

- The female patients in TCGA for LAML have very similar gene expression profiles
- The cell lines are significantly different from the patients and each other
  - One cell line was not substantially more similar to the patient profiles than the other.
- The cell lines may not be a good in vitro model for LAML (based on TCGA patients)

Discussion

- Many details may have effect on result
  - Pair-end vs single-end
  - Strand information
  - Mapping algorithm and RPKM calculation
- Following to do
  - Get more data for cell lines
  - Get data for CML patient
  - Find TF or miRNA regulating differential expressed genes