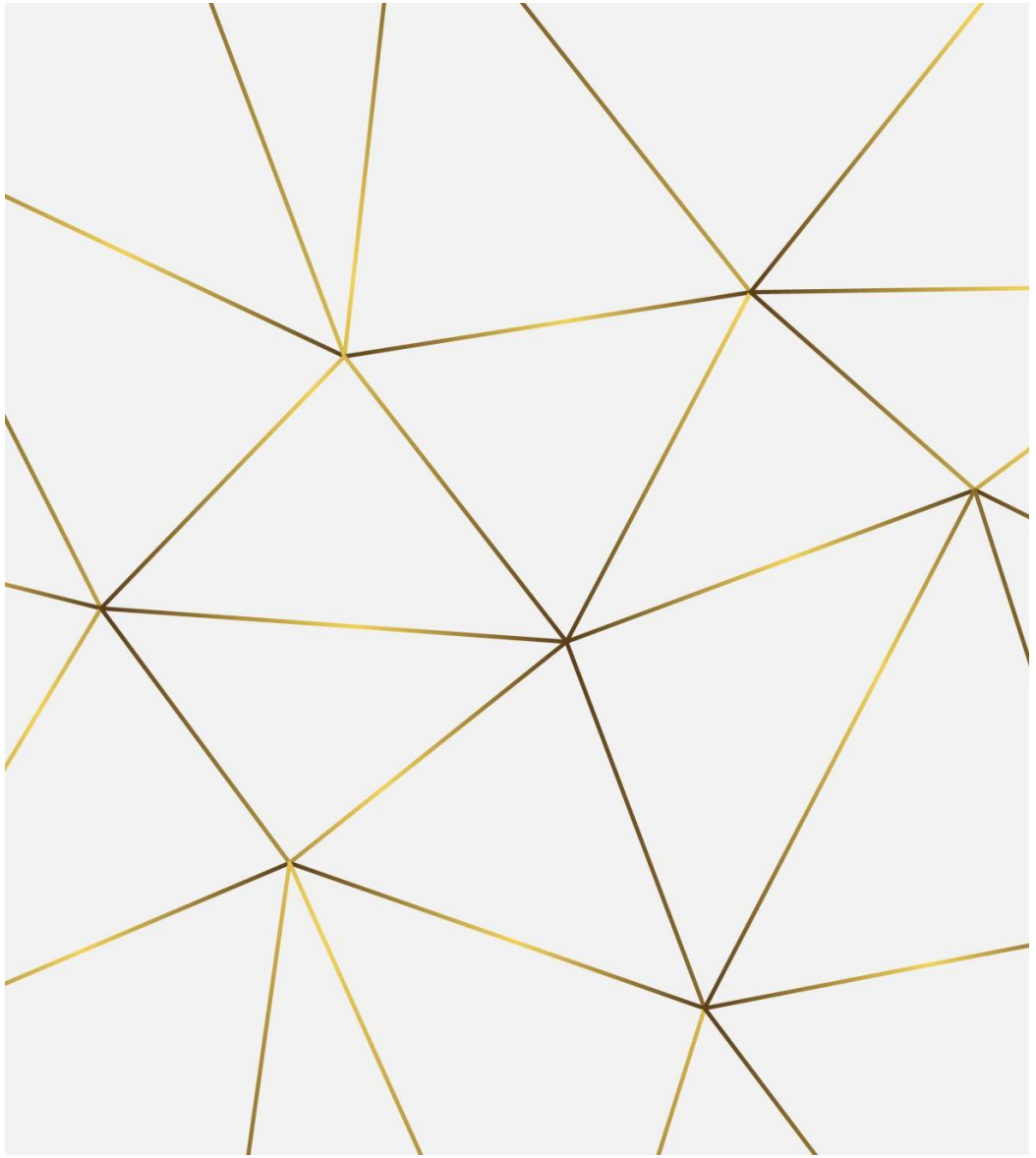
Co-op Project work

Sheetal Nighut, Bioinformatics



Overview

Next Generation Sequencing core @ Elevatebio - July 2022 to December 2022

- MS Bioinformatics student at Northeastern University, Boston

Major projects I worked on

- RNA-Sequencing processing and analysis
- Third-generation sequencing – Pac bio vs Oxford Nanopore
- Amplicon sequencing data transfer via AWS

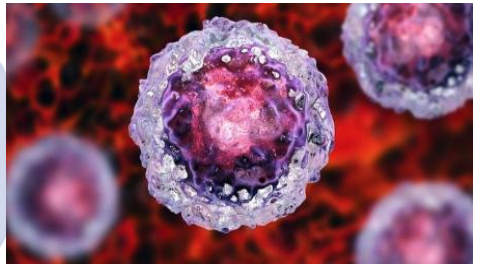
Outcomes

- Boosted teamwork and presentation skills
- Learned to use an analytics dashboard
  - AWS services – EC2 and S3, GitHub
  - R coding - Visualization using ggplot, volcano plot, Heatmap
- Performed RNA-seq pipeline independently on AWS
- Evaluated MultiQC reports, analysis using R studio, and presented the results internally
- Human immune cell types

Background information

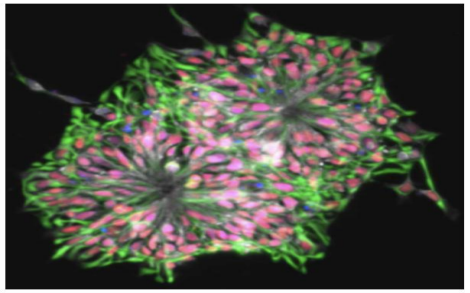
ESC - embryonic stem cell

- Pluripotent stem cells that show an ability to differentiate into every cell type in the human body



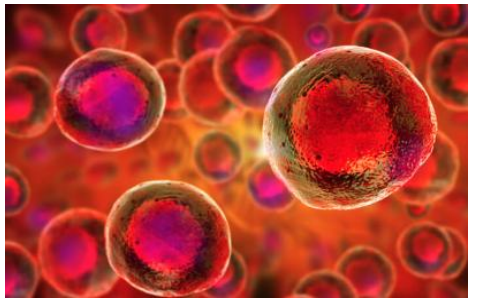
iPSCs - Induced pluripotent stem cells

- Pluripotent stem cells, derived from skin/blood cells
- iPSCs can be differentiated into other cell types including T cells, that may be useful for therapeutic treatments



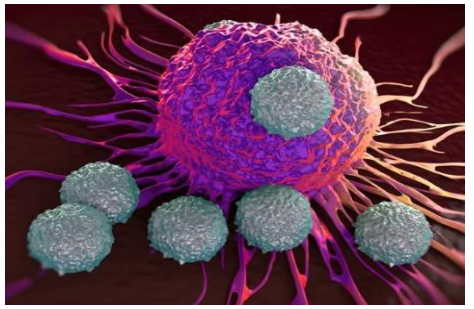
HSPCs - Hematopoietic stem cells

- Stem cell in cord blood that gives rise to blood cell



T cells

- Type of immune cell
- The end goal - make iTcells from iPSCs



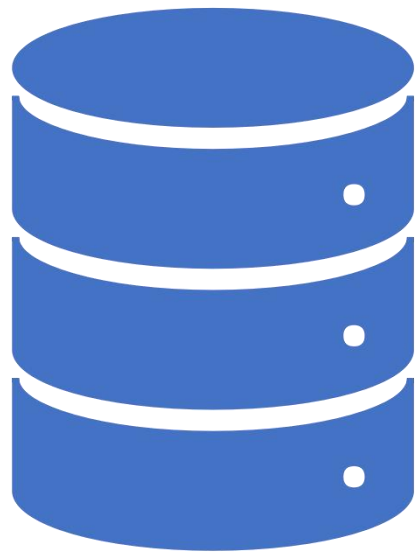
https://www.stemcell.com/hematopoietic-stem-and-progenitor-cells-lp.html
https://www.news-medical.net/health/What-are-T-Cells.aspx
https://stemcell.ucla.edu/induced-pluripotent-stem-cells
https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/embryonic-stem-cell

Transcriptome analysis of early-stage differentiation from iPSCs to iT cells

Outline

1. Experimental Setup
2. Pipeline processing
3. Analysis

1. Experimental Setup



1. Experimental Setup

Goal - Assess early-stage differentiation of iPSCs to iTcell

- 17 RNA-seq samples
- Mix of de-identified samples from background information

Sample and processing

- cell_typeA, cell_typeB, cell_typeC, cell_typeD
- Sorting methods (A, B, C)
- Antigen A fraction (+, -, dim)
- experimental teams – A, B, C

Sequenced on Illumina NextSeq 1000 instrument for 30 hour

Sample_ID
01_cell_typeA_TeamA
02_cell_typeB_TeamB
03_cell_typeB_TeamA
04_cell_typeC_X_antigenA+_TeamB
05_cell_typeC_X_antigenA+_TeamB
06_cell_typeC_X_antigenA+_TeamB
07_cell_typeC_X_antigenA+_TeamC
08_cell_typeC_X_antigenA+_TeamC
09_cell_typeC_X_antigenA+_TeamC
10_cell_typeC_X_antigenA+_TeamA
11_cell_typeC_X_antigenA+_TeamA
12_cell_typeC_X_antigenA–_TeamA
13_cell_typeC_Y_antigenA–_TeamA
14_cell_typeC_Y_antigenAdim_TeamA
15_cell_typeC_Y_antigenA–_TeamA
16_cell_typeC_Z_antigenA+_TeamA
17_cell_typeD_TeamB

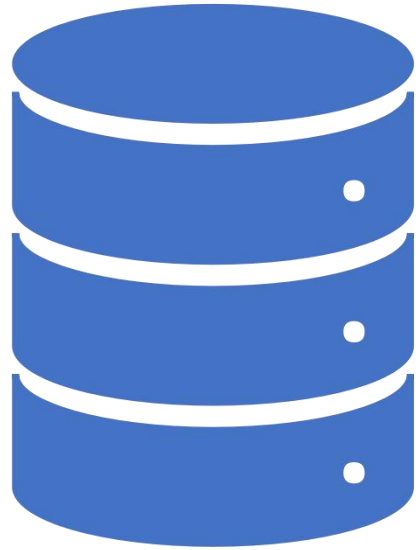
1. Experimental Setup

Analysis objectives :

1) Cell type identity - Cellnet
2) Sample similarity - Compare cell_type C antigen A fraction and sorting methods
3) Differential expression analysis - Compare cell_type C antigen A+ fraction for team B vs team C

2. Pipeline Processing–

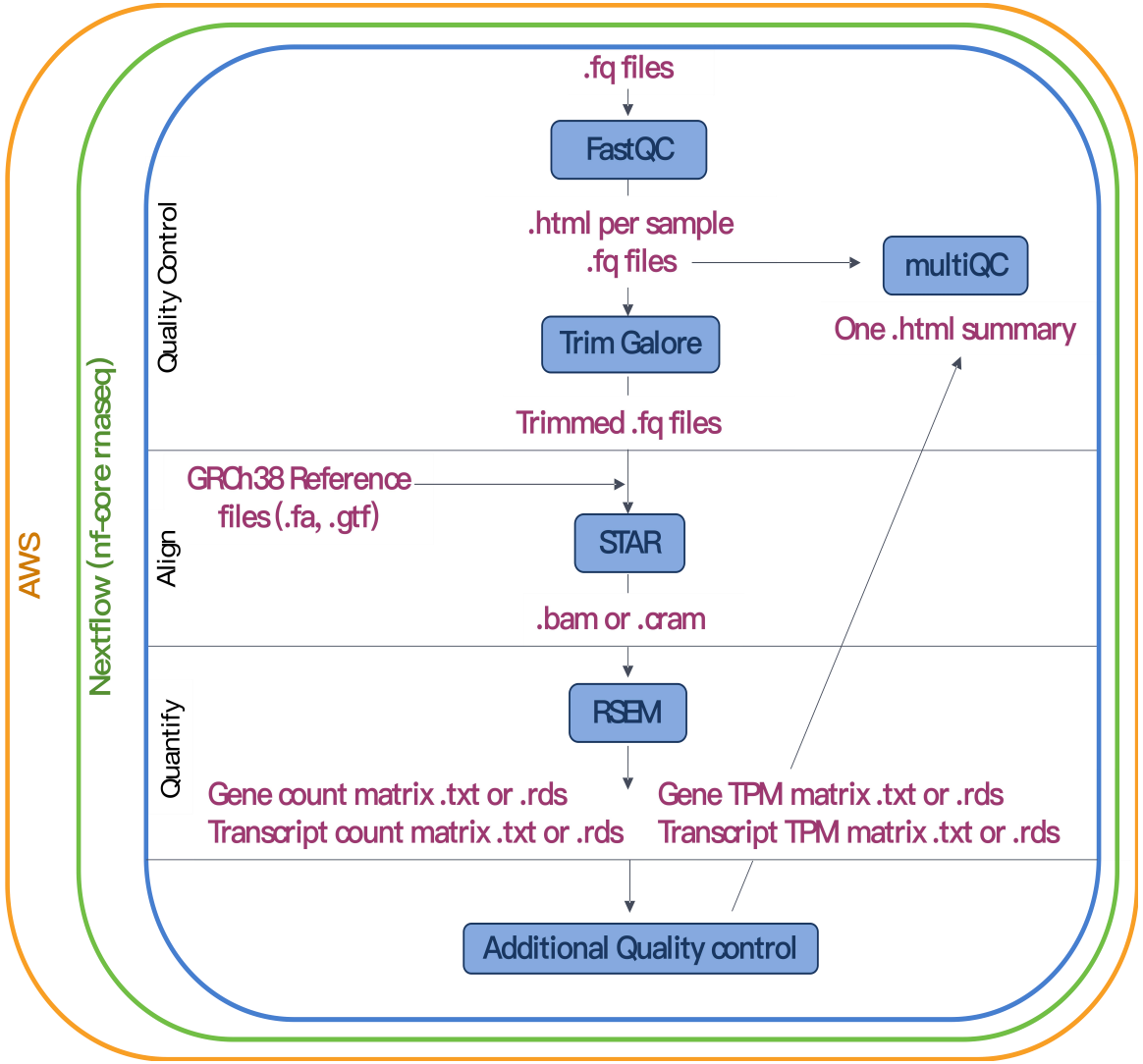
Metrics and QC visuals



2. Pipeline Processing — About pipeline

- Post sequencing, raw data (.fq) are processed through a pipeline in AWS
- Processing gets data analysis ready
  - 30 hr to process (with 488GB mem, 16cpu, 2TB storage)
    - 29972 genes x 17 samples
    - MultiQC report

File
Tool
Pipeline
Workflow manager
Cloud platform



2. Pipeline Processing – Pooled General Statistics

MultiQC

One .html summary

- 17/17 samples with good or useable acceptance criteria
- To meet acceptance criteria, we look at the colored column →
- iPSC - sample 03 shows fewer reads, usable for analysis

Good
Useable
Fail
* two values listed for 1st and 2nd .fq file

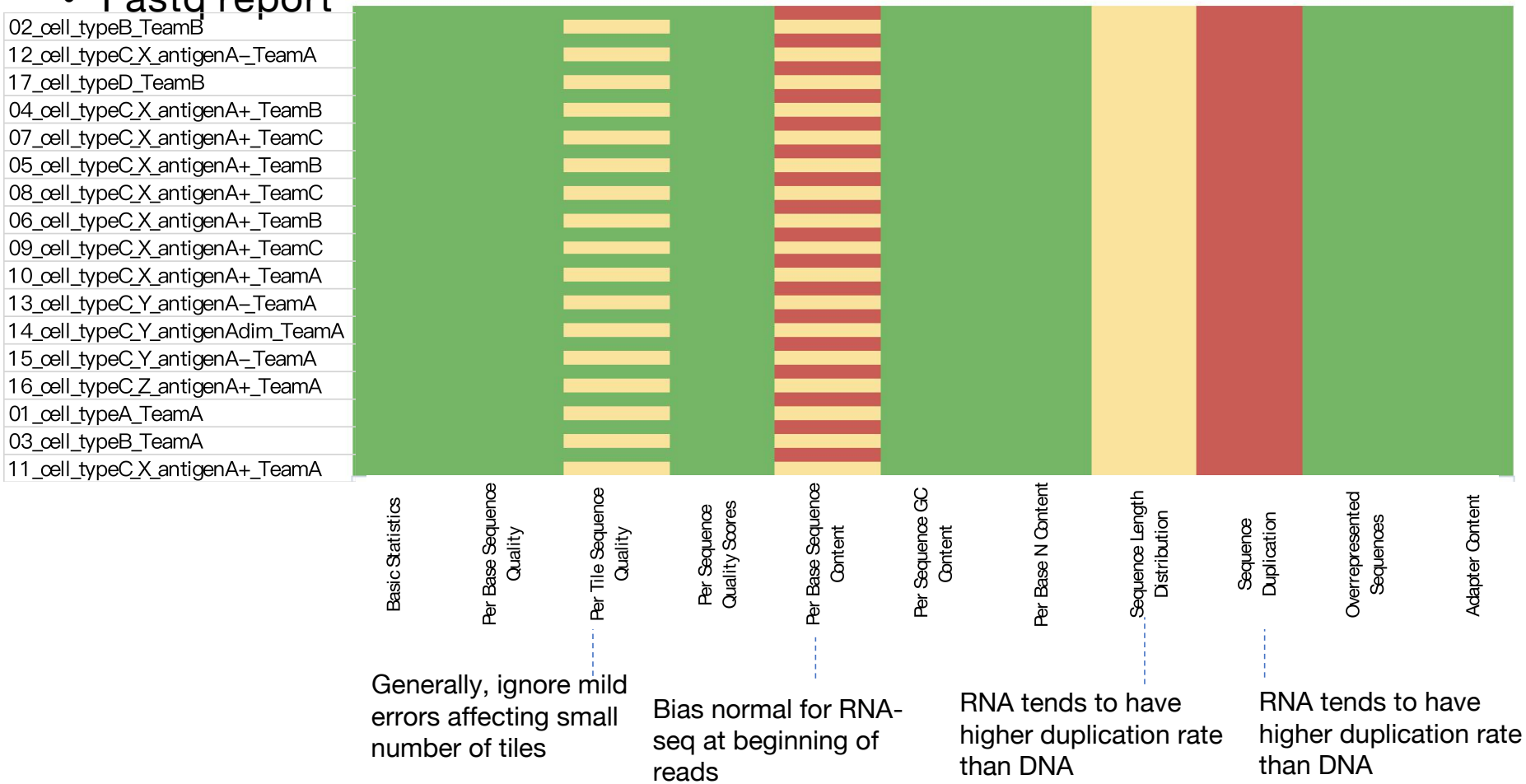
ID		Trimmed bp (%) *	Avg GC content	Avg sequence	Total reads (million) *	Duplication (%)	Alignable reads (%)	Error rate (%)	Reads mapped	Mapped reads (%)	Ribosomal reads (%)
Acceptance Criteria		–	–	–	>20	–	–	–	–	>70	<5
1	01_cell_typeA_TeamA	5.6, 5.6	50, 51	143, 143	29.5, 29.5	34.0	98.5	0.3	58.2	98.8	0.2
2	02_cell_typeB_TeamB	5.0, 5.0	50, 50	144, 144	28.9, 28.9	29.2	98.6	0.3	56.9	98.8	0.4
3	03_cell_typeB_TeamA	7.4, 6.8	50, 50	141, 141	17.3, 17.3	36.9	97.5	0.3	33.7	97.8	0.6
4	04_cell_typeC_X_antigenA+_TeamB	5.1, 5.0	50, 50	144, 143	30.9, 30.9	27.8	97.5	0.3	60.4	98.0	0.3
5	05_cell_typeC_X_antigenA+_TeamB	5.0, 5.0	50, 50	144, 143	26.0, 26.0	25.6	97.9	0.3	51.1	98.3	0.2
6	06_cell_typeC_X_antigenA+_TeamB	5.9, 5.9	51, 51	143, 143	28.2, 28.1	24.8	98.0	0.3	55.2	98.3	0.1
7	07_cell_typeC_X_antigenA+_TeamC	5.7, 5.7	51, 51	143, 142	31.3, 31.3	31.4	97.1	0.3	61.0	97.6	0.3
8	08_cell_typeC_X_antigenA+_TeamC	5.2, 5.2	50, 51	144, 143	30.8, 30.8	26.2	97.4	0.3	60.3	97.9	0.3
9	09_cell_typeC_X_antigenA+_TeamC	5.4, 5.4	51, 51	142, 142	29.3, 29.3	26.6	97.4	0.3	57.2	97.8	0.3
10	10_cell_typeC_X_antigenA+_TeamA	5.5, 5.5	51, 51	143, 143	25.1, 25.1	26.1	98.5	0.3	49.6	98.7	0.4
11	11_cell_typeC_X_antigenA+_TeamA	5.6, 5.5	51, 51	143, 143	25.9, 25.9	29.5	98.1	0.3	50.8	98.3	0.6
12	12_cell_typeC_X_antigenA–_TeamA	5.4, 5.4	51, 51	143, 143	28.9, 28.9	27.2	98.6	0.3	57.0	98.8	0.3
13	13_cell_typeC_Y_antigenA–_TeamA	7.1, 6.9	50, 50	141, 141	33.7, 33.7	25.1	98.2	0.3	66.3	98.5	0.4
14	14_cell_typeC_Y_antigenAdim_TeamA	6.8, 6.6	50, 50	141, 141	28.3, 28.3	32.1	98.3	0.3	55.6	98.6	0.5
15	15_cell_typeC_Y_antigenA–_TeamA	6.0, 5.9	51, 51	142, 142	29.6, 29.6	38.7	98.4	0.3	58.3	98.7	0.3
16	16_cell_typeC_Z_antigenA+_TeamA	6.3, 6.2	51, 51	142, 142	28.4, 28.4	35.3	98.5	0.3	56.0	98.7	0.3
17	17_cell_typeD_TeamB	6.4, 6.3	50, 50	142, 141	29.1, 29.1	23.9	98.1	0.3	57.2	98.5	0.2
Other guidelines		low for Illumina	35-60	–	–	30-90	high	low	–	–	–

2. Pipeline Processing — FastQC Status Check

MultiQC

One .html summary

- Shows status (normal, slightly abnormal, unusual) of .fqs over 11 FastqC metrics
  - Take in context - DNA vs RNA
- Fastq report looks ok for RNA, consistency among samples

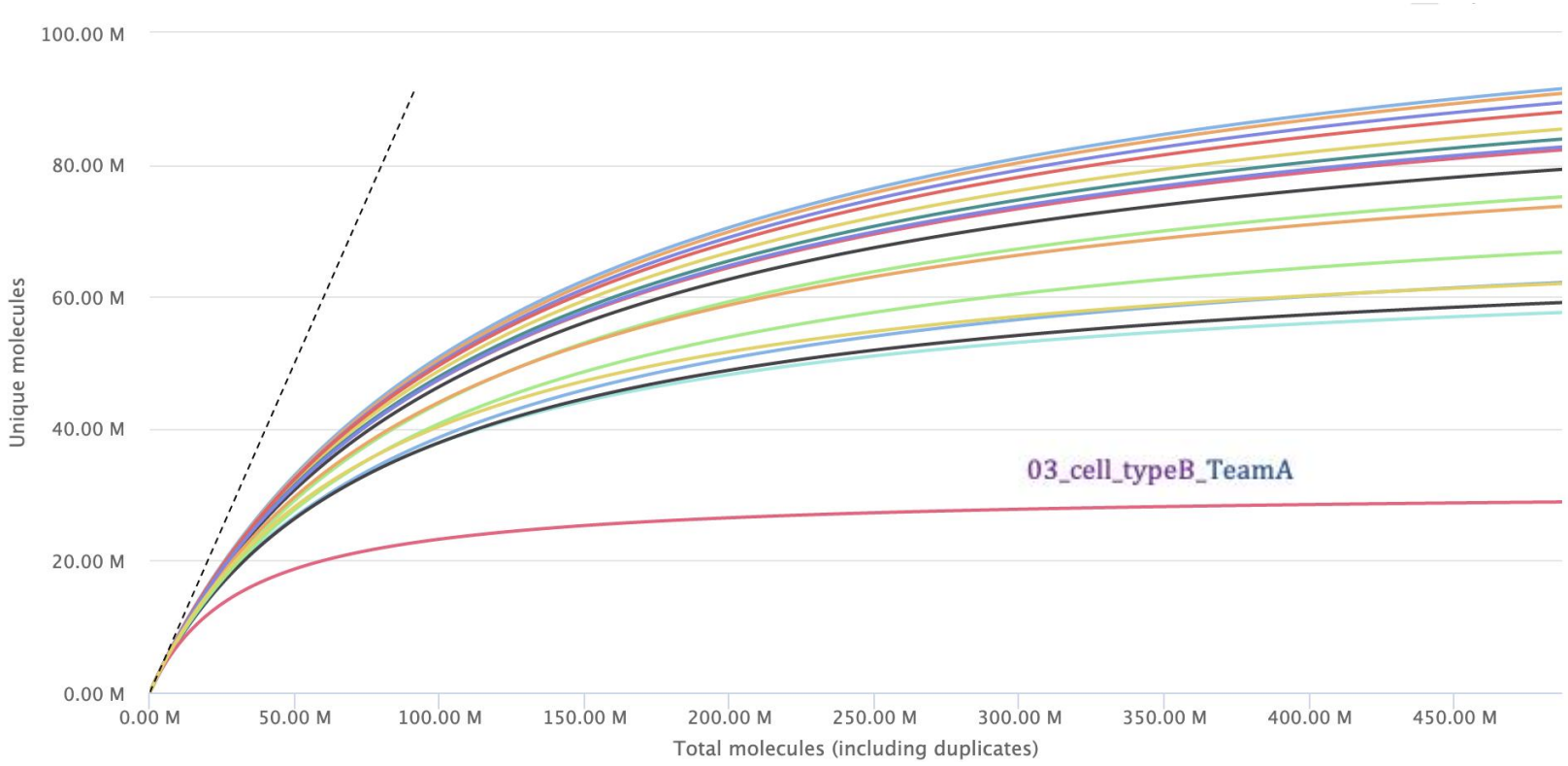


2. Pipeline Processing — Complexity Curve

MultiQC

One .html summary

- Shows total unique reads sequenced as read count increases per sample, → samples have similar complexity saturation trends notable - 03_cell_typeB_TeamA w/ fewer reads
- Different color for each sample, dashed line if every read unique

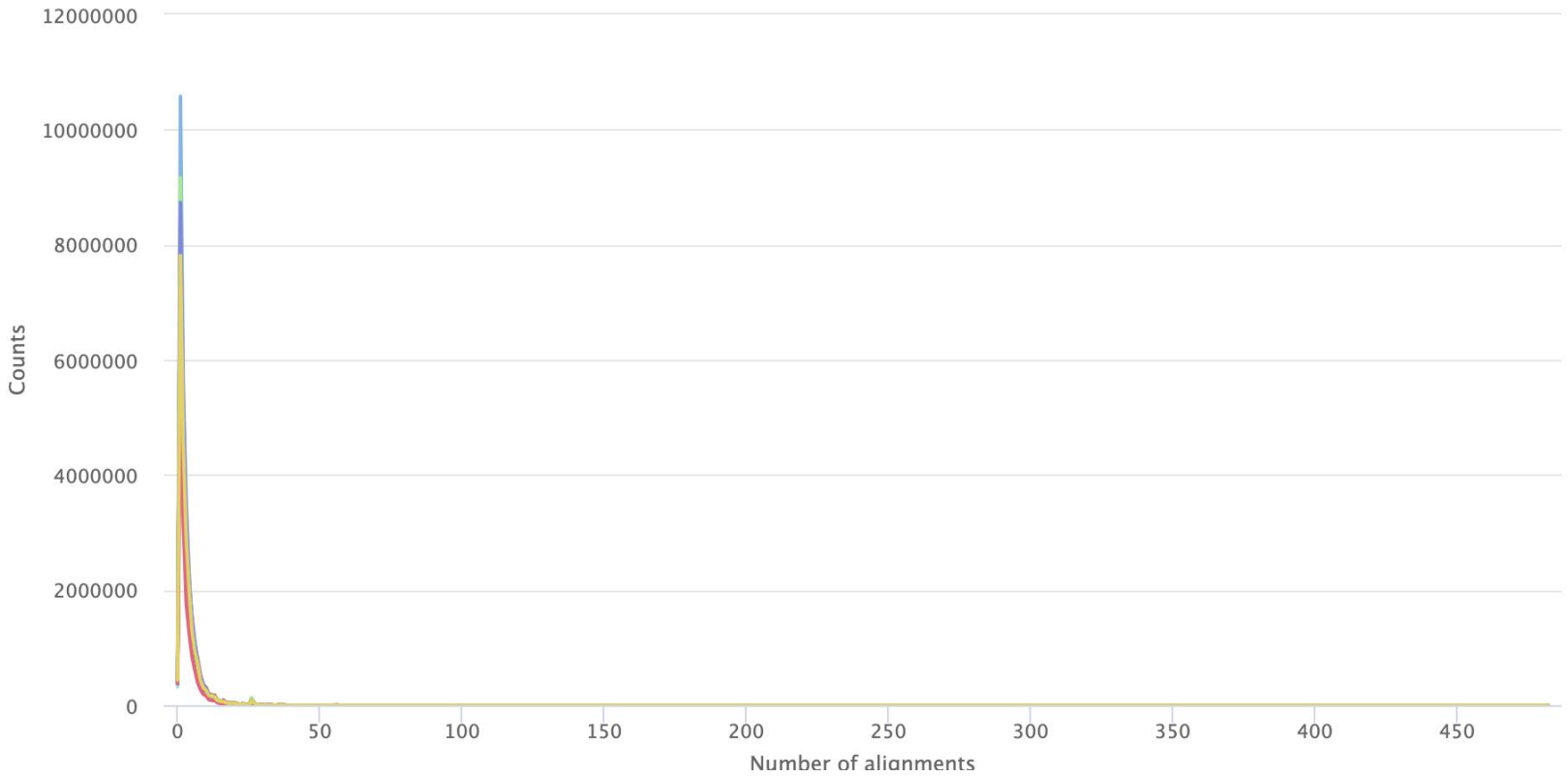


2. Pipeline Processing – Multimapping rates

MultiQC

One .html summary

- Frequency histogram showing how many reads aligned to n reference regions
- Ideal samples should have most reads aligning once (i.e. to a single location) – YES

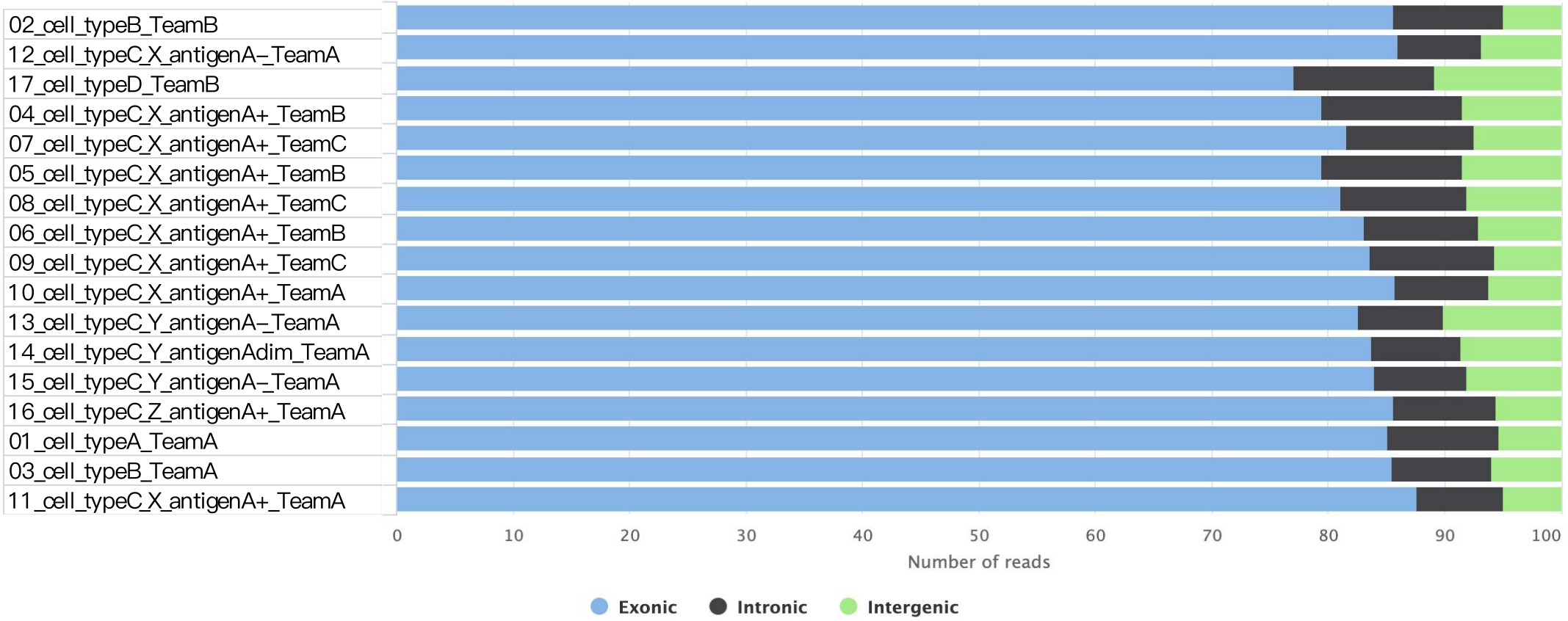


2. Pipeline Processing – Genomic origins of reads

MultiQC

One .html summary

- % of mapped reads classified as originating in exonic, intronic, or intergenic regions
- Isolated mRNA before sequencing, the result matches expectation -mapped reads primarily exons- protein coding region, some introns, and intergenic reads
- Fairly consistent distributions among samples



3. Analysis

3. Analysis — Cell Type Identity (Cellnet)

3. Analysis – Cell Type Identity (Cellnet)

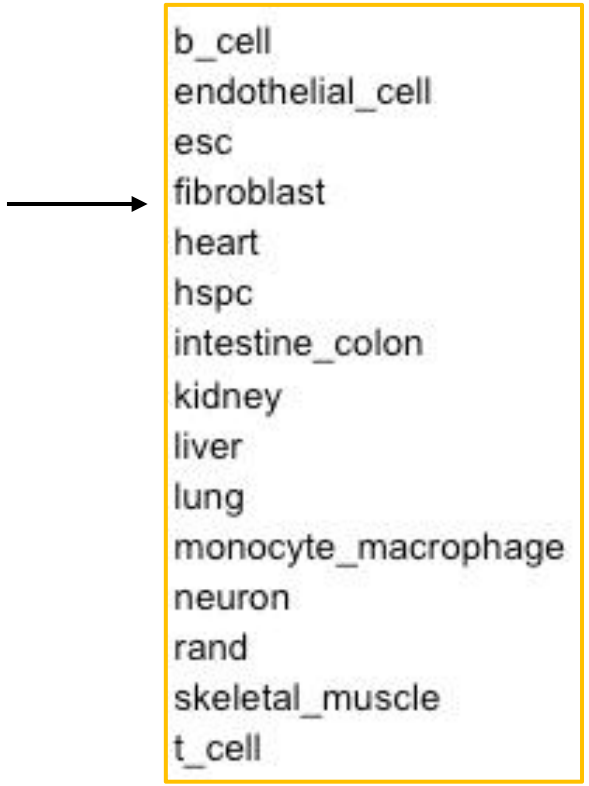
CellNet

- A computational platform that provides cell type identify for 16 cell/tissue types

- Based on gene regulatory networks (GRN) built from public RNA-seq data and a random forest classifier algorithm
  - GRN: a collection of interacting genes with expression trends that govern specific cell types

Output –

- classification score, heatmap



CellNet: Network Biology Applied to Stem Cell Engineering, https://doi.org/10.1016/j.cell.2014.07.020

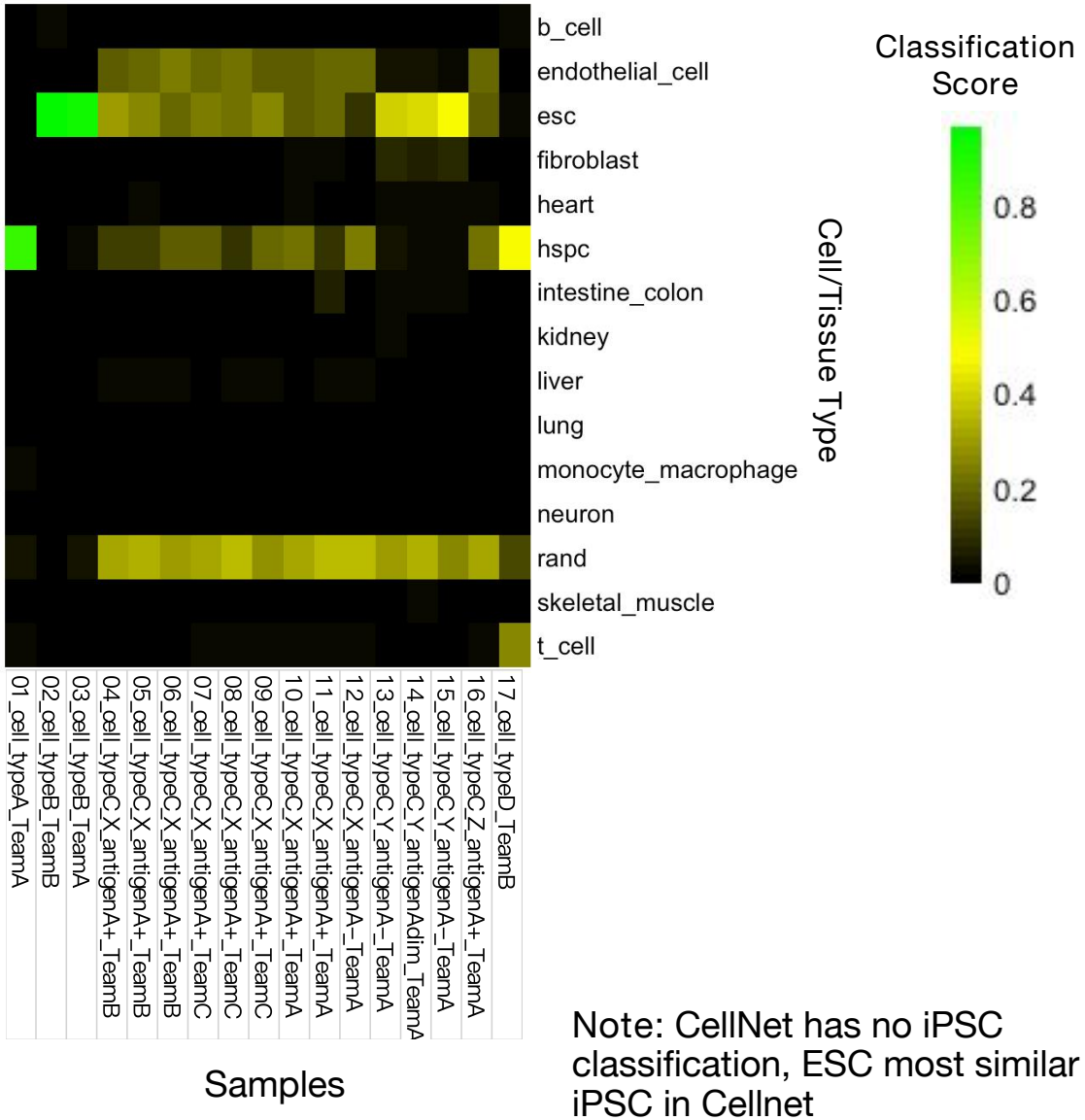
HSPC - hematopoietic stem and progenitor cells
ESC – Embryonic Stem Cell

Cell Type Identity CellNet: Classification Heatmap

0 = unlike training tissue
1 = indistinguishable from training tissue

Results

- Cell_typeA sample 1 shows a high score for HSPC
- Cell_typeB samples 2-3 show high scores for ESC
- Cell_typeC samples 4-16 show varied low scores for random, HSPC, ESC, and endothelial
  - Cell_typeC with sorting Y samples 13-15 (antigen A- and dim) have higher ESC score
- Cell_typeD sample 17 shows a moderate score for HSPC, a low T cell score
- CellNet classification matched expectations for all samples



Note: CellNet has no iPSC classification, ESC most similar to iPSC in Cellnet

3. Analysis – Sample Similarity (DESeq2)

3. Analysis – Sample Similarity (DESeq2)

PCA (PC1 Vs PC2)

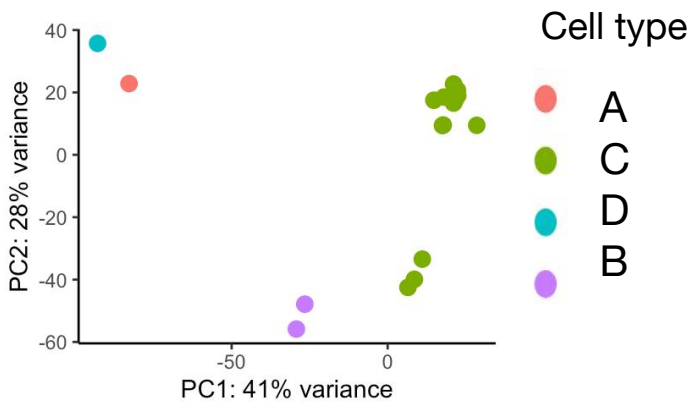
- PC1 41% explained variance
- PC2 28% explained variance

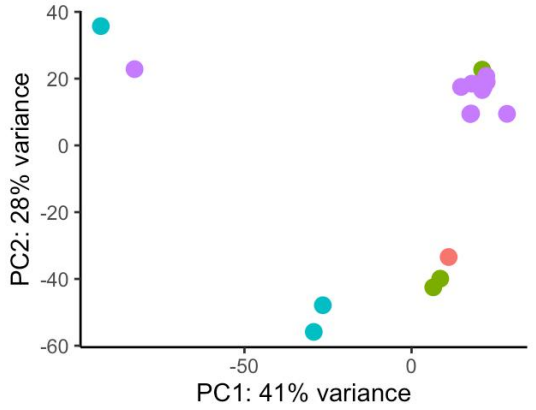
Results

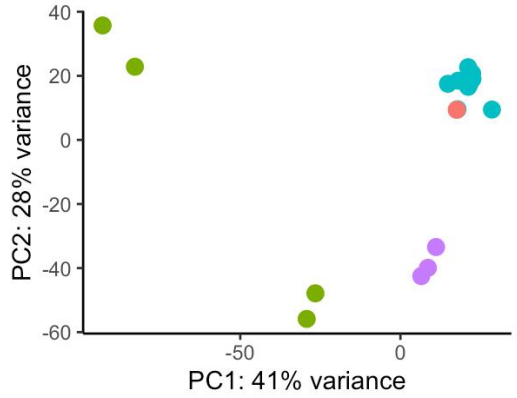
Cluster by cell type

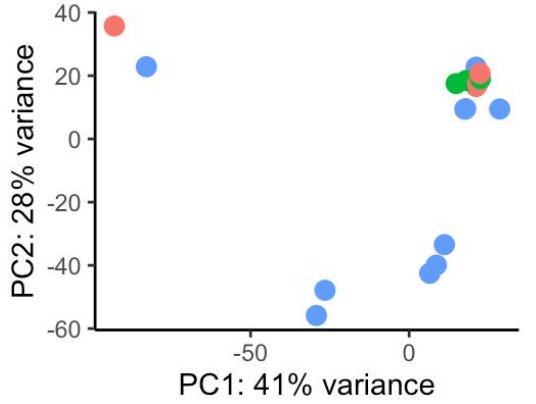
- Cell_typeB, Cell_typeA near cell_typeD, cell_typeC
- Y sorting cell_typeC separate from other sorting methods- X and Z
- Cell_typeC antigen A + seemingly cluster, few neg and dim cell_typeC samples – the trend could be confounded by sorting method

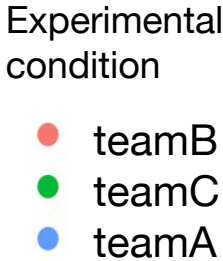
X Antigen A+ Cell_typeC samples close in PCA space regardless of experimental condition (team A, B, or C)











3. Analysis – Sample Similarity (DESeq2)

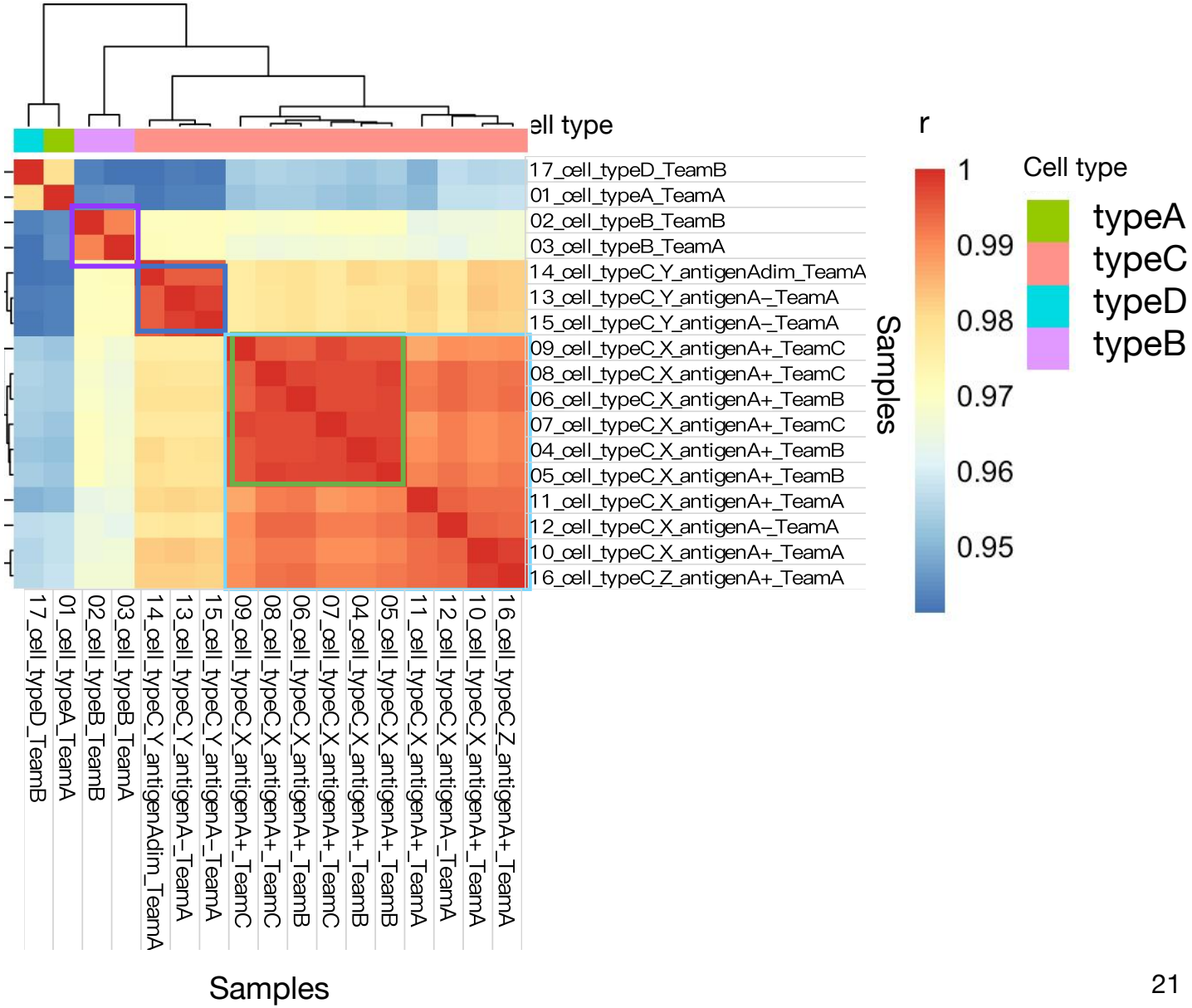
Heatmap of sample-to-sample Pearson correlation (r)

- Hierarchal clustering of rlog values, heatmap colored by r value

Results

Cluster and high correlation by cell type

- Cell_typeB, cell_typeA near cell_typeD, cell_typeC
- Special notes - high correlation among:
  - typeB antigen A+ samples for team B and team C
  - Y sorting typeC antigen A negative and dim samples
- typeB and typeD, are most different from other samples
- Trends support PCA results



3. Analysis – Team B vs C antigen A+ typeC (DEA, DESeq2)

3. Analysis – Team B vs C antigen A+ typeC (DEA, DESeq2)

Differential expression analysis

6 of 17 samples from the same controlled experiment →

- typeC antigen A+, X sorted samples from team B OR team C
- Saw High similarity in the Pearson heatmap and PCA

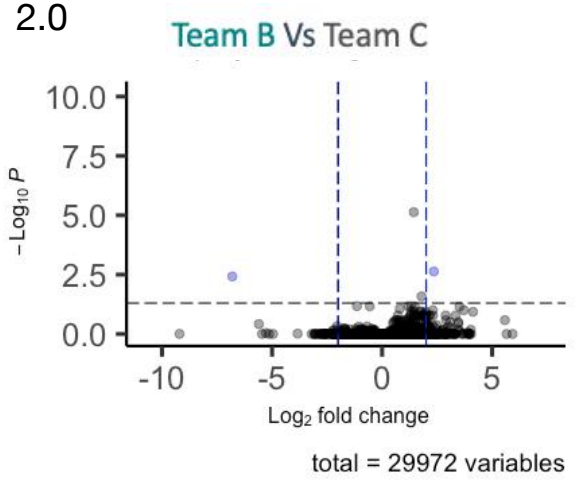
04_cell_typeC_X_antigenA+_TeamB
05_cell_typeC_X_antigenA+_TeamB
06_cell_typeC_X_antigenA+_TeamB
07_cell_typeC_X_antigenA+_TeamC
08_cell_typeC_X_antigenA+_TeamC
09_cell_typeC_X_antigenA+_TeamC

- DEA table

Genes	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
A1BG	283.159699	-0.251056026	0.18479796	-1.3585433	0.17429136	0.99950086
A1BG-AS1	78.5995045	0.230925629	0.32700915	0.70617484	0.48007942	0.99950086
A1CF	5.16622568	1.684699231	1.13095323	1.48962768	0.13632216	0.99950086
A2M	9217.26931	0.59809265	0.75001859	0.79743711	0.42519722	0.99950086
A2M-AS1	78.134544	0.181366767	0.33852803	0.53575111	0.59213057	0.99950086
A2ML1	11.0725648	0.901692975	0.79593209	1.13287677	0.25726601	0.99950086

- Assess DEA results with volcano plots
  - logFC by –log adj pvalue
  - Each point represents one gene (n=29972)
    - colored = DEG, black = not

Cutoffs adj pvalue < 0.05, logFC ⩾ 2.0



2 total DEGs

- Result
  - Show high similarity with only 2 differentially expressed genes
  - To gain insight into a mix of de-identified samples starting early-stage differentiation of iPSCs to iTcell, would require more samples

Conclusion

- Successfully processed and analyzed transcriptome for 17 RNA-seq samples
  - 17 samples passed QC checks
  - mix of typeA, typeB, typeC, typeD, sorting methods, antigen A fractions

- Cell type identify – typeC samples with no high classification scores, low score mix of ESC, HSPC, random, and endothelial cell. Cell typeA and typeB samples met expectations

- Sample similarity - expected cluster and correlation by cell type
  - typeC sorting by Y separate from X and Z typeC samples with confounding factors such as antigen A+ fraction
  - antigen A+ fraction typeC samples from teams B and C show high similarity

- Differential expression analysis - same controlled experiment, only 2 differentially expressed genes, again, high similarity

- To gain insight into a mix of de-identified samples starting early-stage differentiation of iPSCs to iTcell, would require more samples

Acknowledgments: work is achieved with team collaborations ☺

Cherylene

NGS Core

Cellular Engineering

Adriana

Will

Christina

Sheetal

Gary

Kelsey

Kyle

Seth