Solving Biological Problems that require Math 2016

Investigating gene expression across the cell cycle using single-cell RNA-seq data

Adam Alexander T. SMITH
post-doc in ACM lab, UNIL

UNIL | Université de Lausanne

SIB

Biological Background:
lncRNAs & cell cycle

Long non-coding RNAs: the new guys in town

Bulk RNA-seq revealed extensive, non-protein-coding transcription of genomes

ncRNAs are diverse
micro RNAs, piRNAs, pseudogenes... and lncRNAs

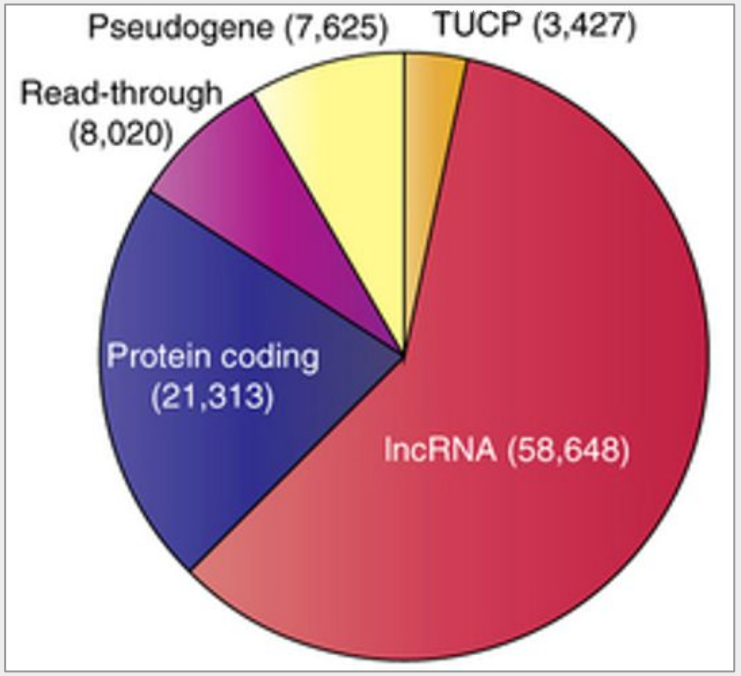
lncRNAs: ~60% of known Hu transcripts, functions discovered for a few 100's

e.g. HOTAIR, lnc-DC, Xist...

Expression: High tissue specificity

Diverse regulatory roles, often involved in cancer.

A whole new layer of gene expression regulation!!!

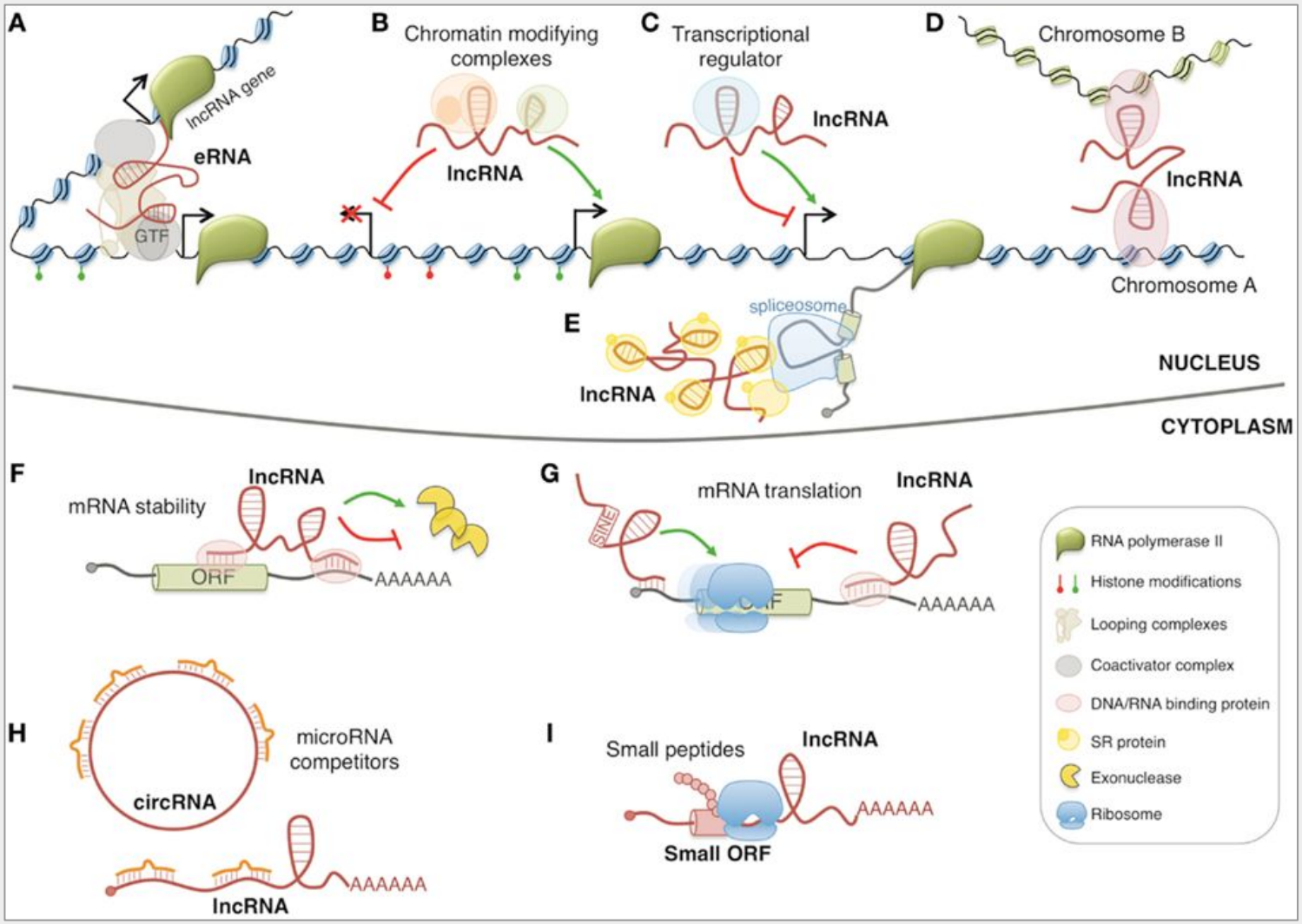


Number of transcripts in the human genome: a large fraction of transcriptional information is in non-coding regions
Lyer et al., 2015

lncRNA definition & characteristics - or lack thereof

- >200nt
- Lack an (obvious) open reading frame
- Pol II transcribed, often spliced, capped, poly-adenylated like mRNAs
- Lower evolutionary conservation than mRNAs
- Little known about domains, 2^{ry} structure, etc
- No widely-accepted lncRNA sub-classification
  Current best: relative to genomic protein-coding regions
  (anti-sense, intronic, intergenic…)
  => ACM lab focuses on *intergenic* lncRNAs
- Generally lower expressed, expression highly time- & tissue-specific

Various regulatory lncRNA roles



Morlando et al., Front. Med. 2015

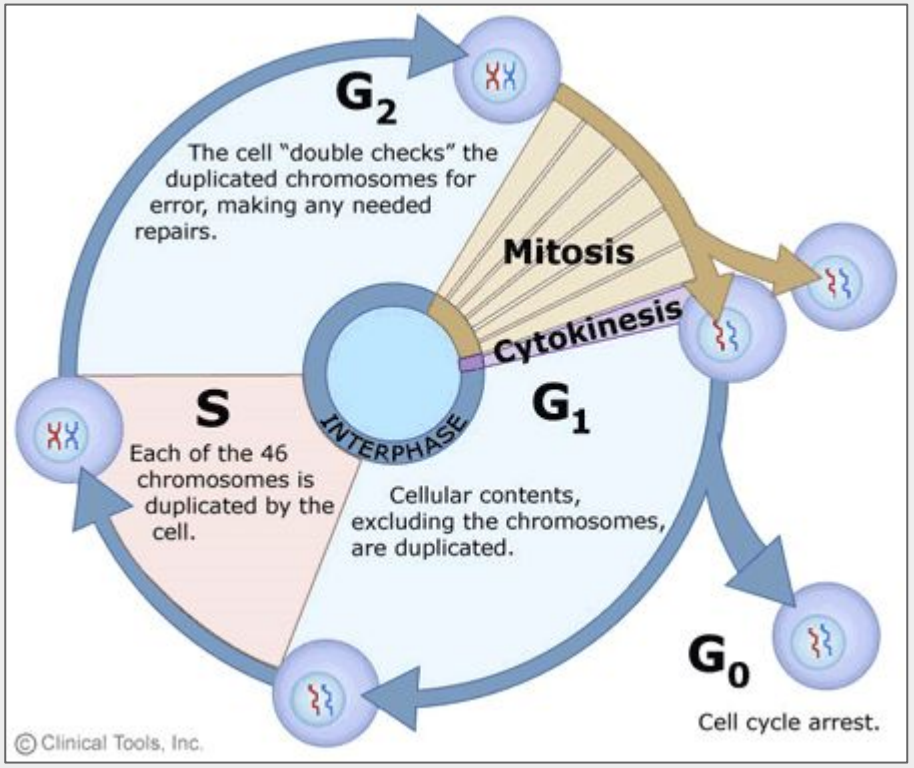
2nd aspect: cell cycle

Cell cycle

Extensively studied:

- gene & protein expression
- phenotypes
- yeast, bacteria, mammalian cells

BUT:
Not so well-studied from a ncRNA viewpoint!



Work Hypothesis: lncRNAs are involved in the cell cycle

Why would it make sense for lncRNAs to be involved in cell cycle?

- High tissue & time specificity
  => role in tightly-controlled regulatory processes
- low expression
  => responsive processes
- RNA-based mechanisms faster than protein-based ones
  => responsive processes
- Some annotated lncRNAs shown to be involved in the cell cycle
- Recent work @ACM lab: lncRNAs regulating mRNA via miRNAs: mRNAs enriched in cell cycle genes

New experimental techniques

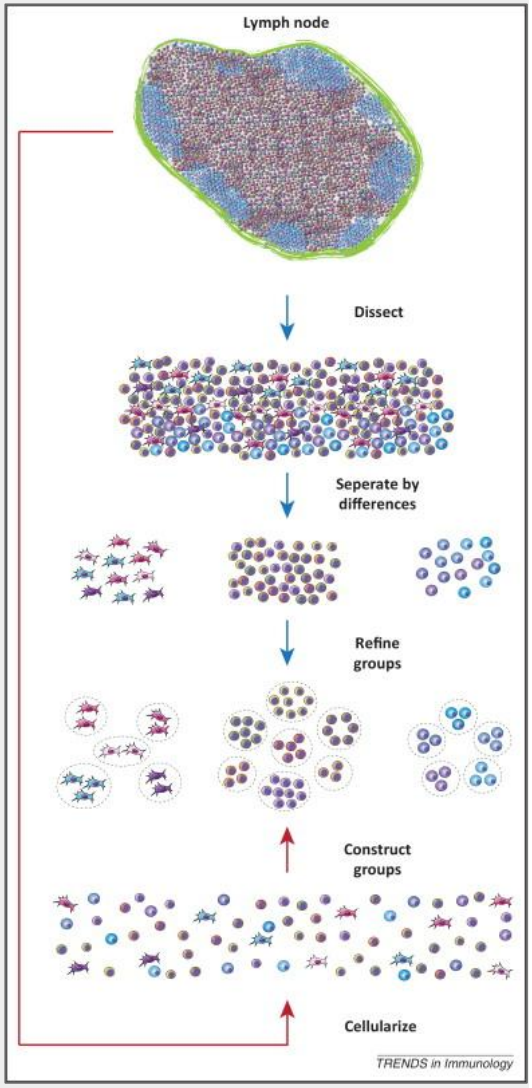
Methodology Background: scRNAseq

The advent of single-cell RNA sequencing

RNA-seq has come a long way…
sequences RNA from 100's of 1000's of cells
=> assuming sample cell population is homogenous,
could only see strong, cell population-average signals

scRNA-seq has recently emerged
hundreds of cells, relatively shallow sequencing
=> sample cell heterogeneity can be explored!

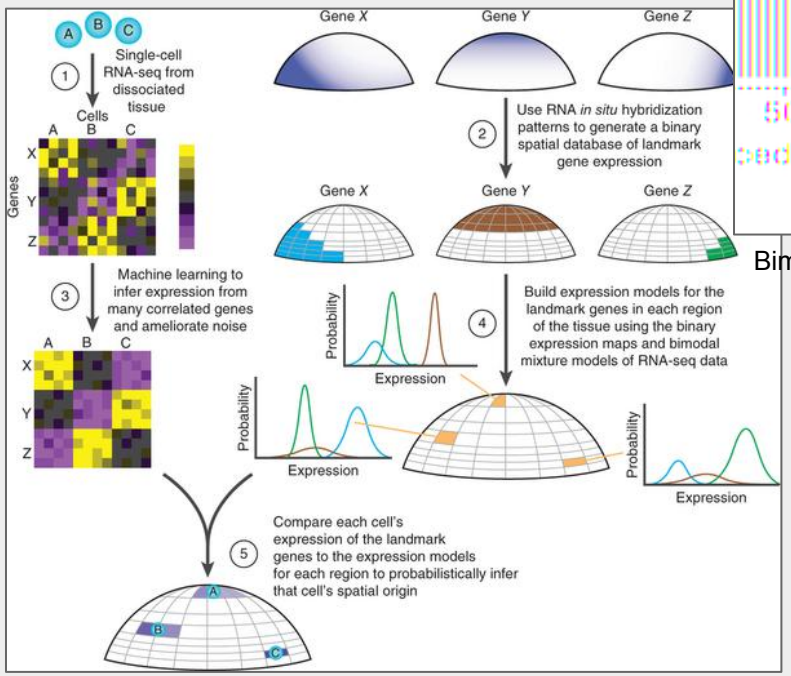
- sub-populations
- differentiation & response dynamics
- cell cycle

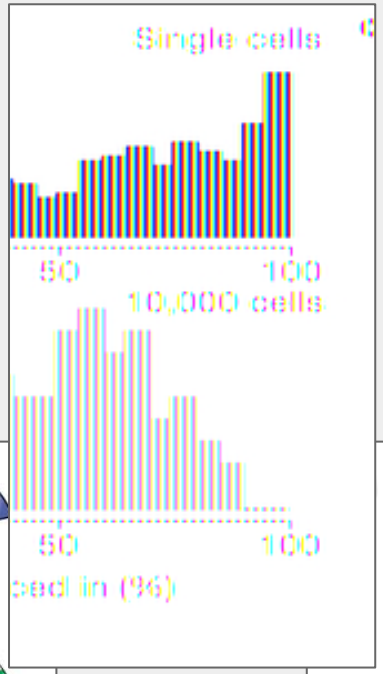


Satija & Shalek (2014)
Trends Immunol.

scRNAseq application examples

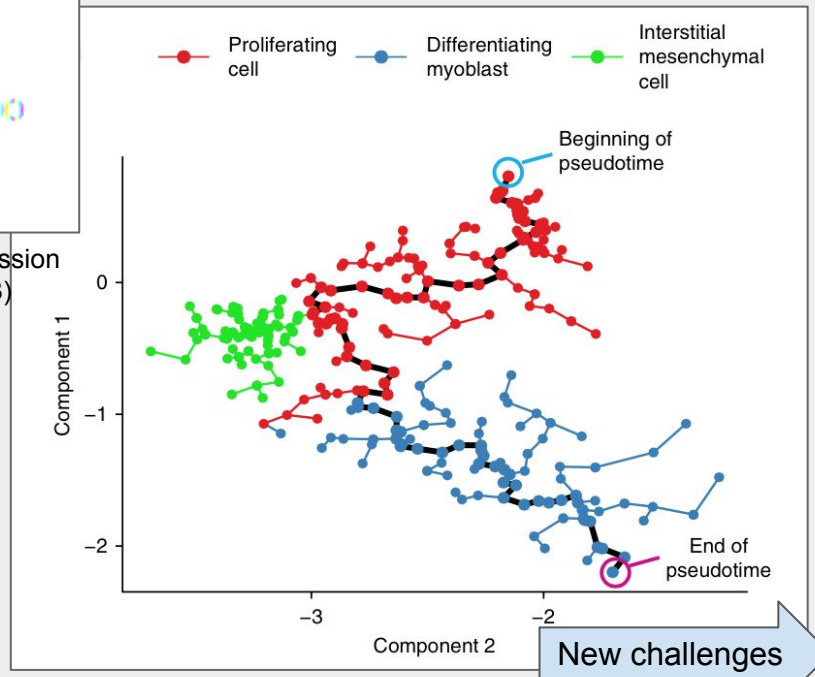
Spatial Mapping of single cells
Satija et al. (2015)
Nat Biotech





Bimodal isoform expression
Shalek et al. (2013)
Nature

Differentiation path reconstruction
Trapnell et al. (2014)
Nature



New challenges

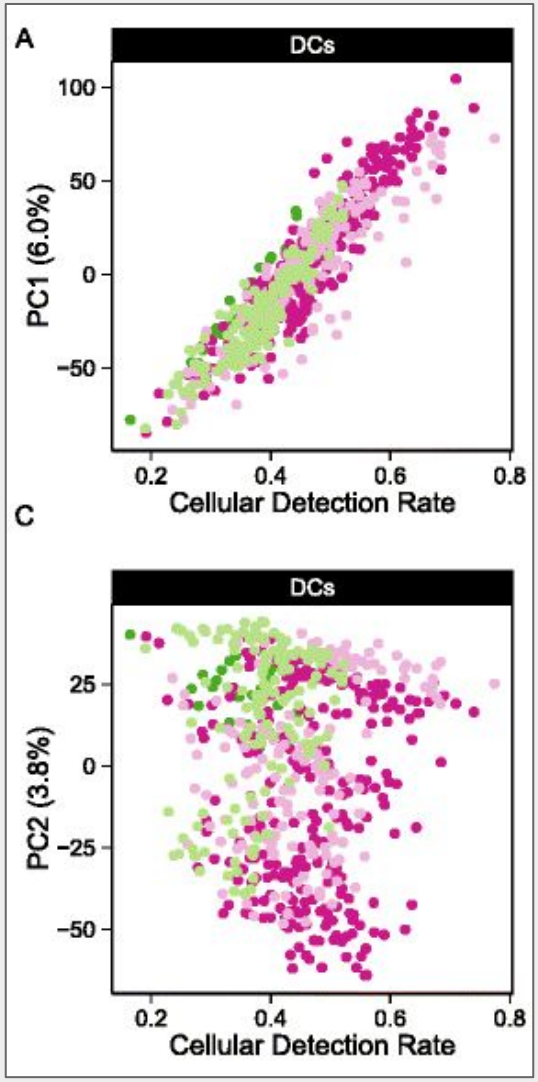
“Cellular detection rate”: a major signal in scRNAseq

Transcript Cellular Detection Rate (CDR):
%cells in which the transcript was detected

Biological biases:
cell volume, bimodal expression, stochasticity

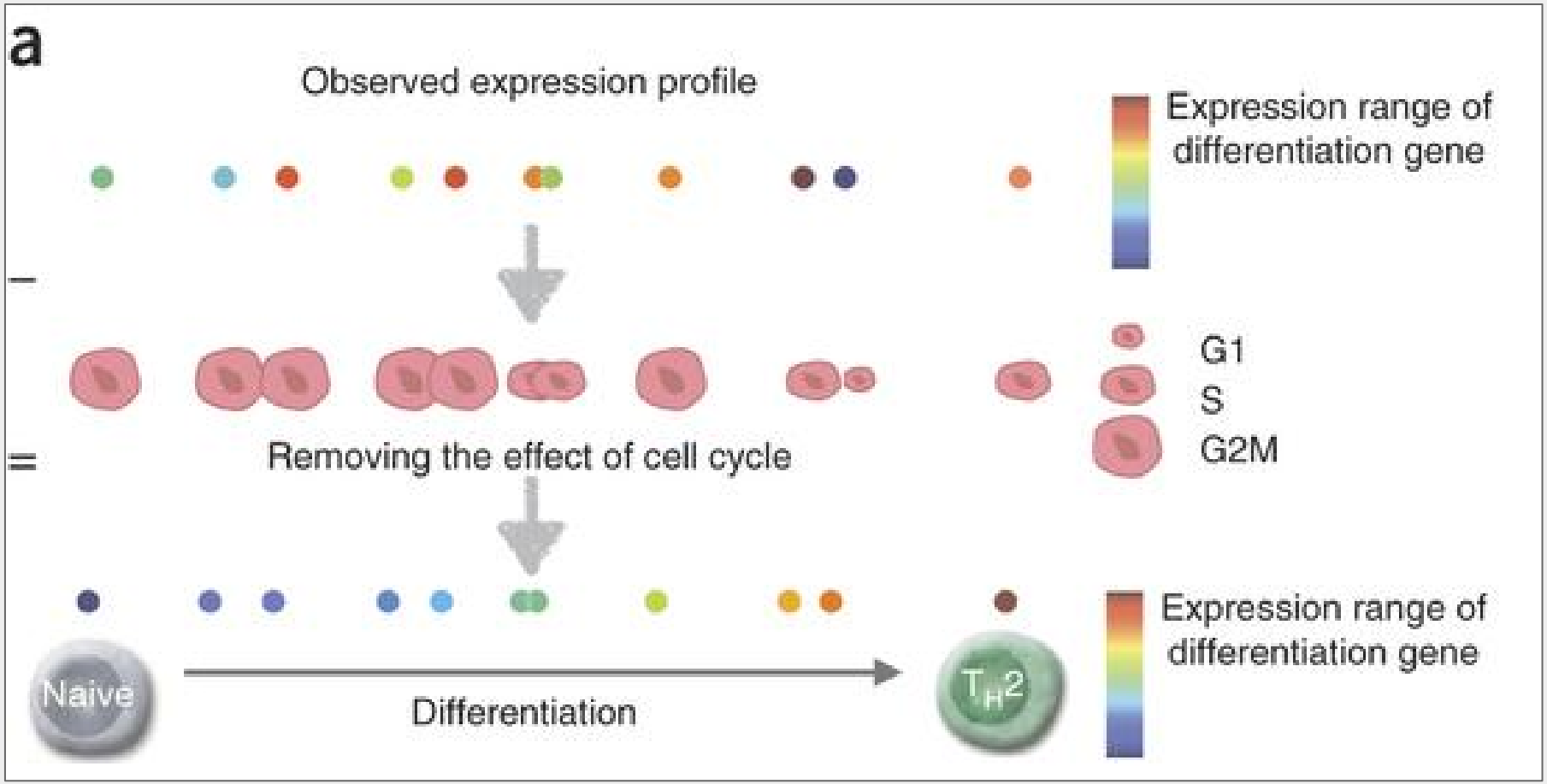
Technical biases:
low RNA qty, amplification, shallow sequencing

=> frequent non-detections of genes/transcripts, esp. for lowly expressed
=> major signal in scRNAseq



Finak et al. (2015) Gen. Biol.

Cell cycle: a major confounder in scRNAseq



Cell cycle: a common nuisance to remove. But what if you want to study it?

New methods

Technical Background: pseudo-alignment

Recently developed “pseudo-alignment” transcript quantification methods

Rather than locating the exact alignment of an RNA-seq read on the Genome/Transcriptome, find known transcripts whose k-mer “summaries” are the most compatible with those of the read
=> MUCH faster (1000x!)

Kallisto
Bray, N., Pimentel, H., Melsted, P., and Pachter, L. (2015).
Near-optimal RNA-Seq quantification with Kallisto. arXiv:1505.02710 [cs, Q-Bio].

Sailfish
Patro, R., Mount, S.M., and Kingsford, C. (2014).
Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotech *32*, 462–464.

Salmon
Patro, R., Duggal, G., and Kingsford, C. (2015).
Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. bioRxiv 021592.

The Bootstrap Benefit

With such fast methods, can BOOTSTRAP the data

Bootstrapping (statistics): random re-sampling with replacement

Here: randomly re-sample RNA-seq reads, with replacement
similar to what RNA-seq does already IRT RNA in the samples

=> Generate pseudo-technical replicates

So what?
=> can evaluate robustness of transcript expression estimates
=> can use in downstream statistics

Current Work

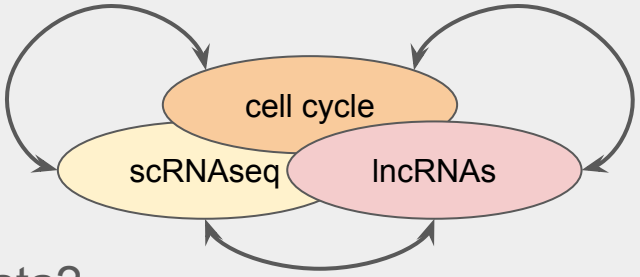
Bioinformatics Project @ ACM

Analysis of publicly-available **scRNA-seq data** on **staged mES cells** to identify candidate **lncRNAs** expressed & differentially expressed at various stages of the cell cycle.

Biological Question:
can cell cycle signal be analysed (rather than removed) from scRNAseq data?

Technical Question:
can lncRNAs be detected in scRNAseq data, given their generally low expression?

Biological question:
are lncRNAs involved in the cell cycle?



Methods Question:
can mapping-free methods be applied to scRNAseq data?

Dataset

Publicly-available single-cell RNA-seq dataset [E-MTAB-2805]

96*3 Mouse Embryonic Stem Cell samples
staged for G1, S, G2/M based on Hoechst 33342 staining (FACS area)

6.5-7.5 M paired-end reads / sample
ERCC spike-ins

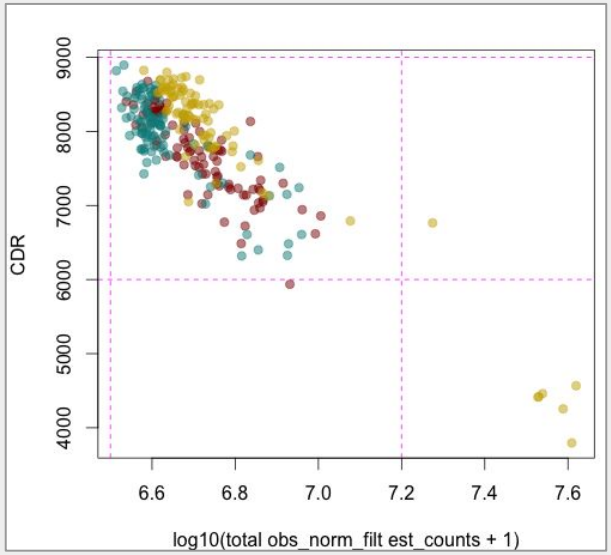
Publicly-available ENSEMBL mm10 genome

With genome annotation: 111,709 transcripts
+ in-lab selection of 9,757 lncRNA transcripts

Buettner, F., et al. (2015). *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells*. Nat Biotech 33, 155–160.

Methods

- Transcript expression levels (estimated counts) quantified using **Kallisto**, a command-line tool
- Expression levels analysed with **Sleuth**, a companion R package to Kallisto
- Single-cell data **QC** in R, inspired by original paper
- **PCA**s in R for exploratory analyses

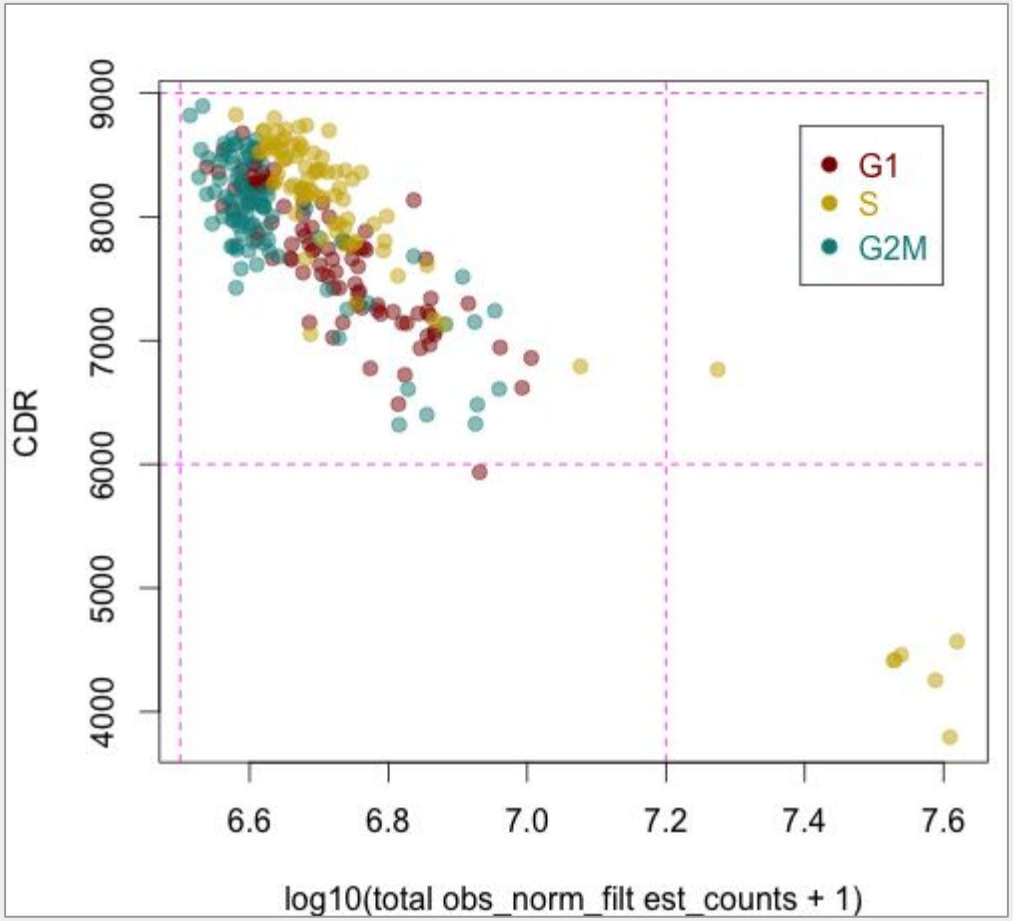


One scRNAseq QC plot

scRNAseq QC

Multiple measures were considered during QC of single cell data:

- Sleuth filter:
  9,611 / 121,466 transcripts
- Total number of hit counts in [$10^{6.5}$; $10^{7.2}$]
- Number of genes detected in [6k; 9k]

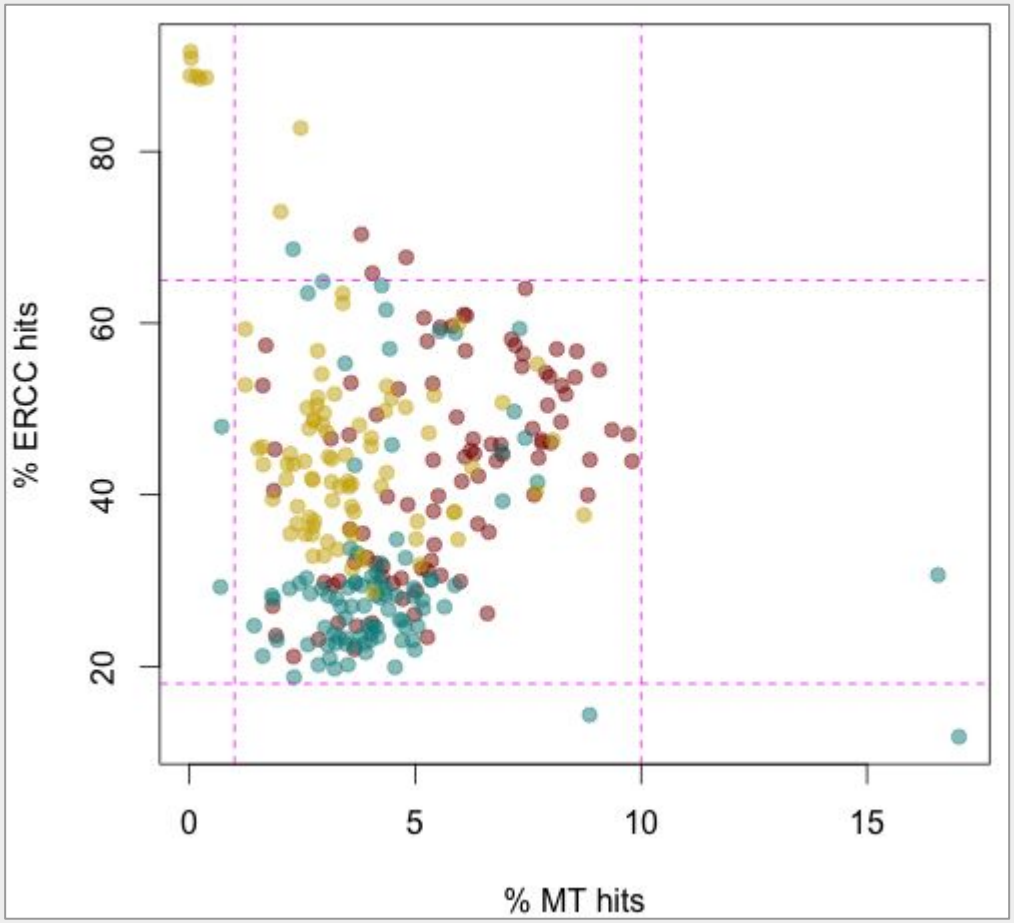


scRNAseq QC

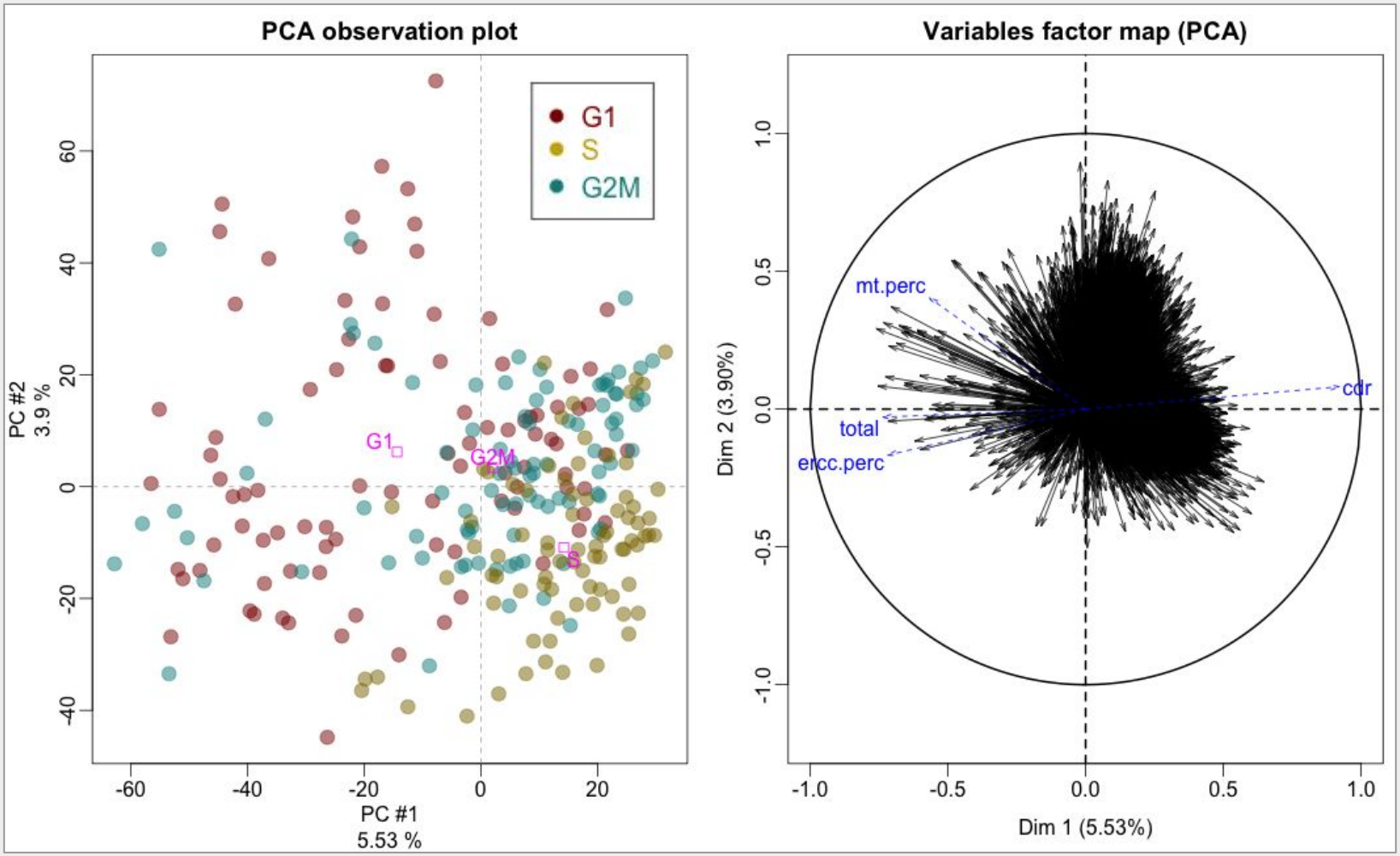
Multiple measures were considered during QC of single cell data:

- Sleuth filter:
  9,611 / 121,466 transcripts
- Total number of hit counts
  in [$10^{6.5}$; $10^{7.2}$]
- Number of genes detected
  in [6k; 9k]
- Percentage of hits to MT
  in [1%; 10%]
- Percentage of hits to ERCCs
  in [18%; 65%]

=> 92 G1, 80 S, 90 G2/M cells

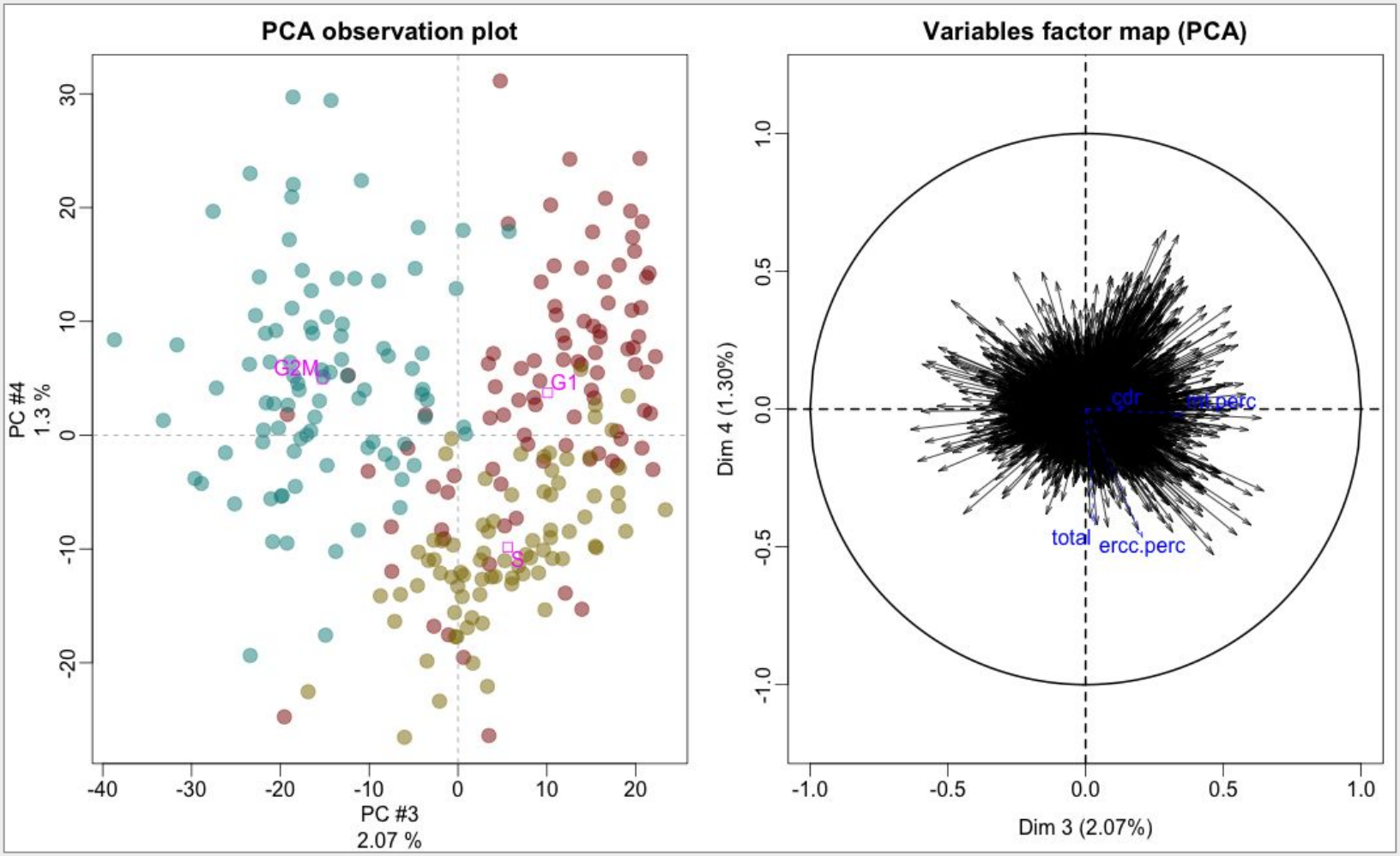


Overall PCA



=> Predominant **CDR signal**… what about on the next PCs?

Overall PCA



=> Clear cell cycle signal

What will YOU do?

a selection of the following...

Analysing cell cycle stage DETs

If requested, possibility of processing data from the command line
All further analyses will be carried out in R (likely via RStudio)

Answering Biological Questions...

- Identify DETs between cell cycle stages
- Determine DET specificity
  - Are DETs stage-specific? Anti-specific? Shared?
- Determine lncRNA distribution amongst DETs
  - Broad Institute Gene Set Enrichment Analysis
- Intersect results with biological knowledge
  - Enrichment Analyses: Pathways, GO terms, MGI phenotypes...

In red: priorities

What does the bootstrapping bring to the final analysis?

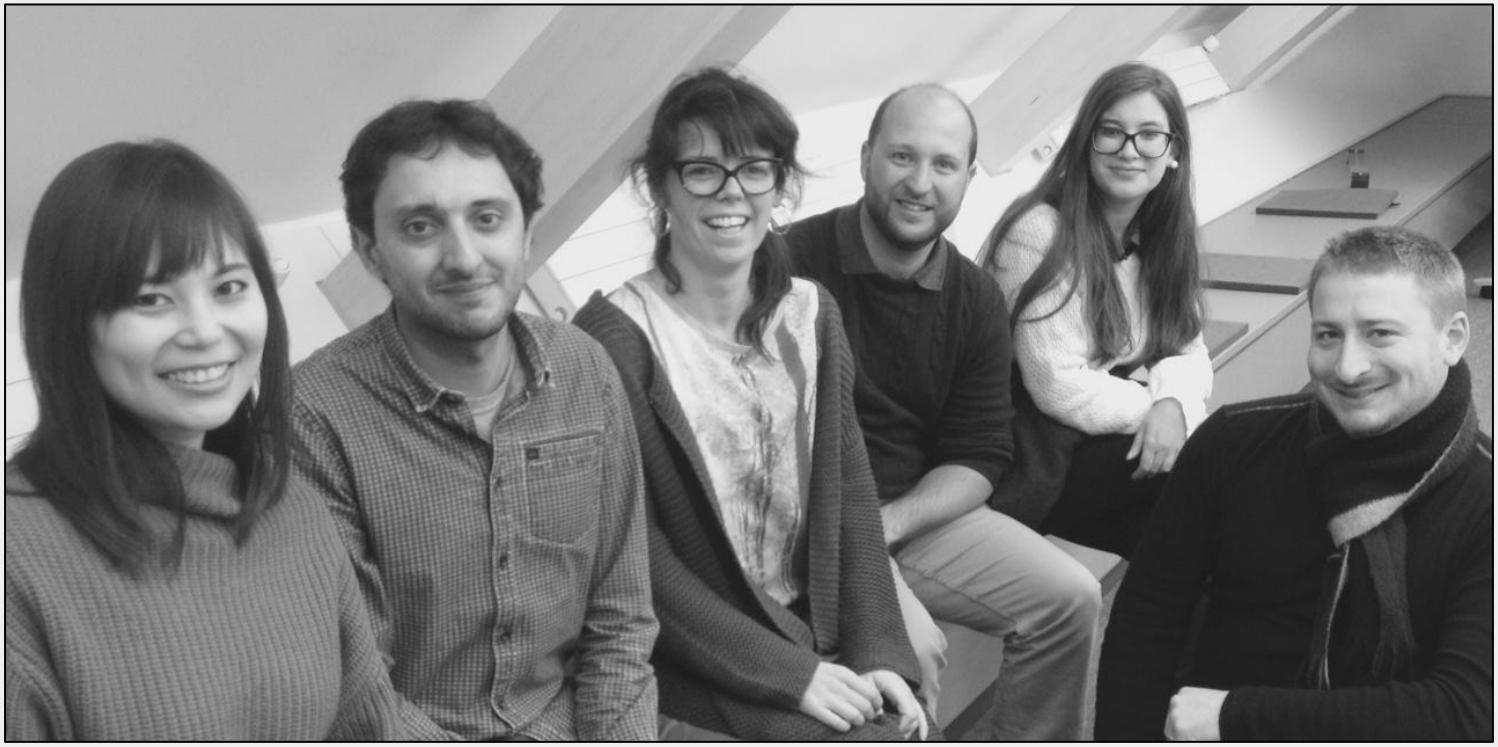
Answering Technical Questions...

- Determine if the bootstrapping helps inform the scRNAseq QC
  - effect on CDR, hit rates...
- Understand Sleuth's filtering IRT the bootstrapping
  - reverse-engineering or empirical

Answering Statistical Questions...

- Try using the bootstrapped samples be used in further statistical analyses
  - do they add to the interpretation of PCAs?
  - can they be used as technical replicates?

In red: priorities



We look forward to seeing you!

Thank you for your attention