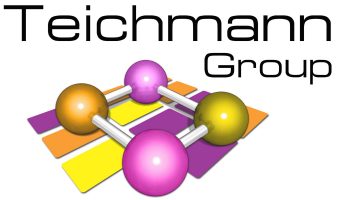
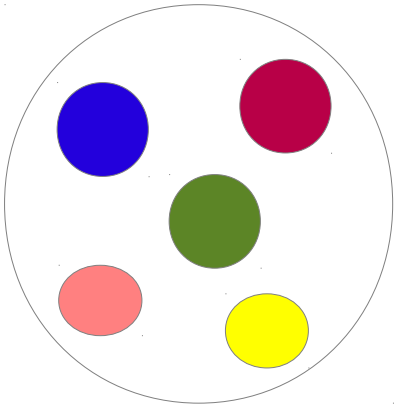
Scoring transcript variation in single cell RNA-seq data

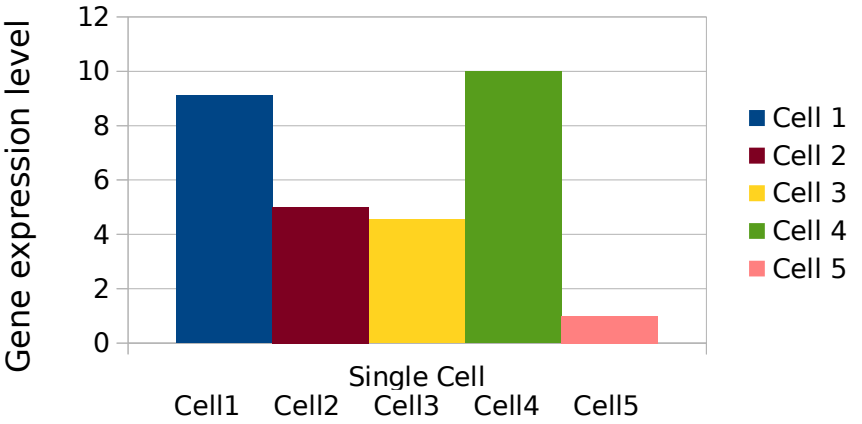
Xiuwei Zhang



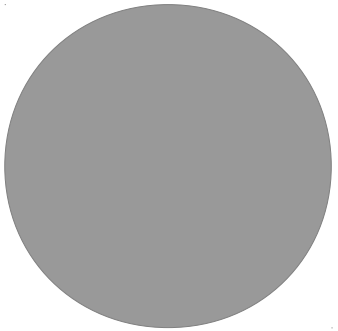
Single cell RNA-seq provides data at cellular resolution

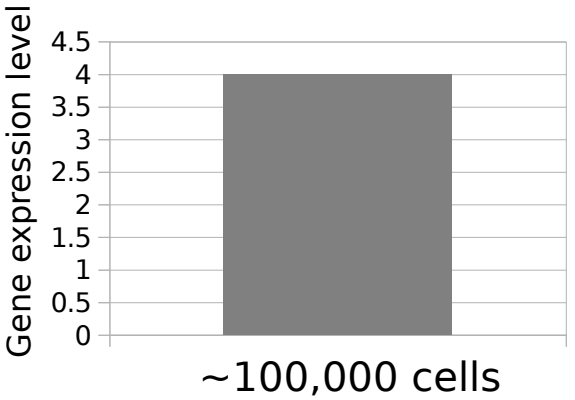
Single-cell RNA-Seq



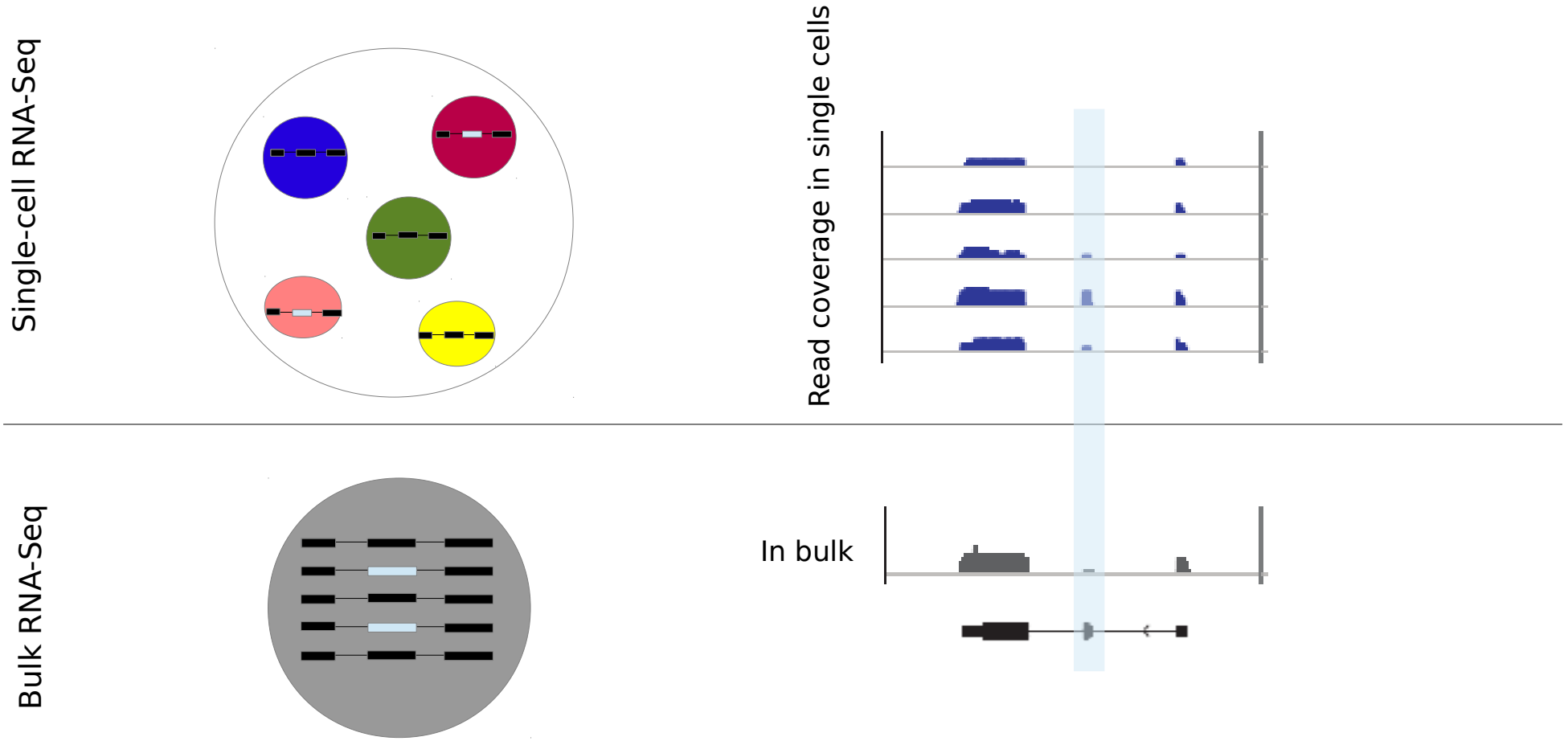


Bulk RNA-Seq





Single cell RNA-seq provides data at cellular resolution



Single cell RNA-seq also shows variation in read coverage profiles

Background

Traditional bulk RNA-seq tools for single cell data

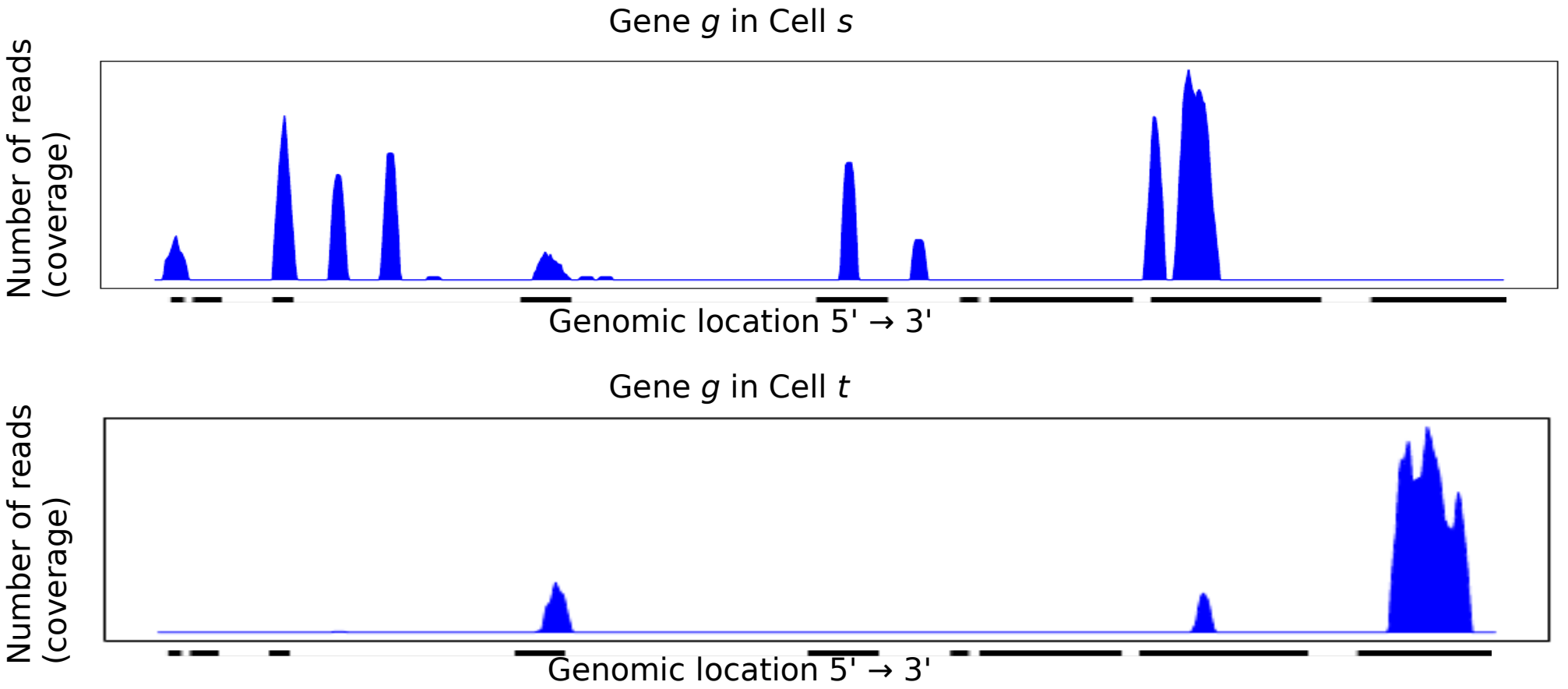
- Calculating exon/intron inclusion scores:
  - MISO (Katz *et al*. 2010)
    used in Shalek *et al.* 2013
    high isoform variation, bimodal distribution of PSI scores
  - Bam2ssj (Pervouchine *et al*. 2013)
    used in Marinov *et al*, 2014
    number of isoforms for one gene in a cell
- Find novel splice junctions

Global analysis of profile variation across single cells
Sources? patterns? sub-populations?

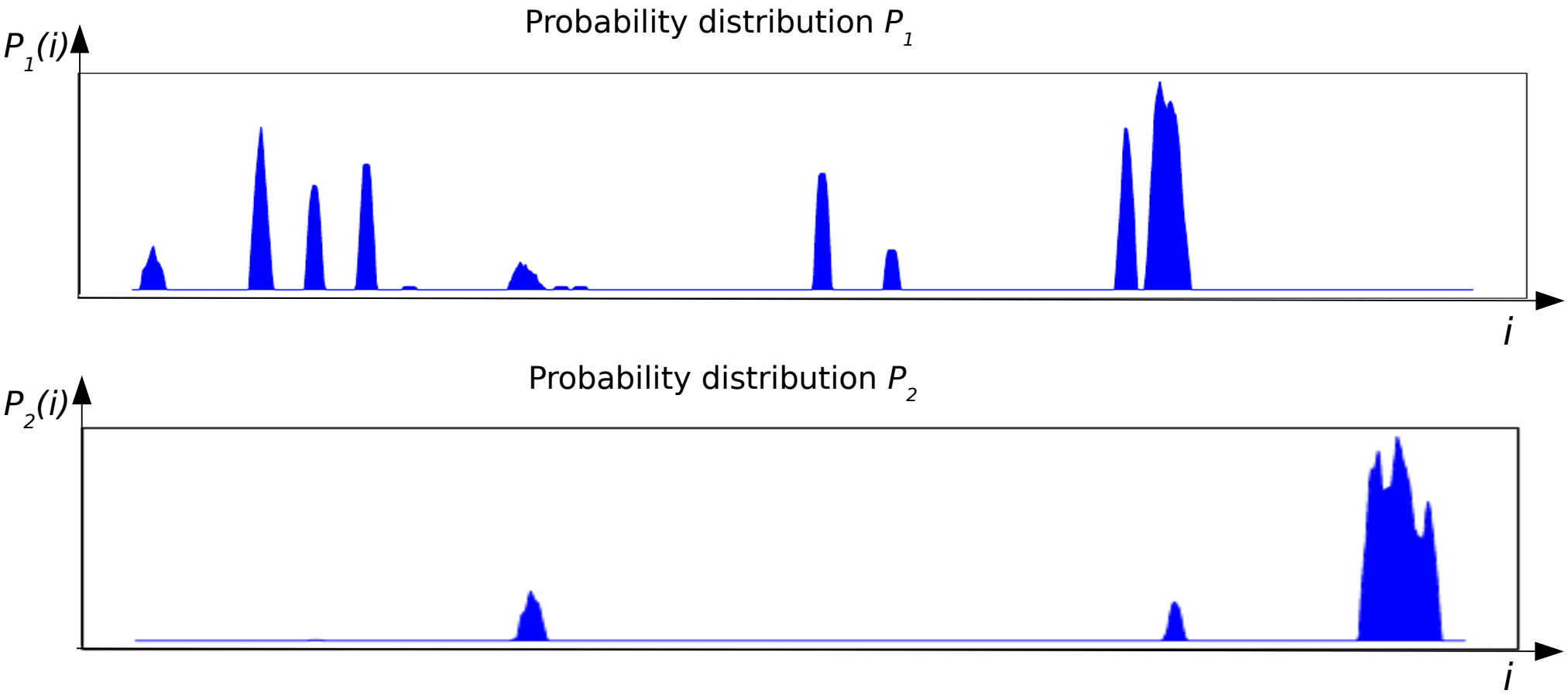
Outline

- Method
  - Profile Variation (PV) score
- Benchmarking and thresholding
  - Various data sets
  - Various gene categories and exons
  - Compare with bulk RNA-seq
- Applications
  - Genes with high isoform variation
  - Patterns in isoform usage
  - Genes which switch isoforms

Profile variation (PV) score



Profile variation (PV) score



Jensen-Shannon Divergence (JSD)

Profile variation (PV) score

$$\mathrm{JSD}(P_1, P_2, \ldots, P_n) = H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i)$$

H( ): entropy

increases with the number of categories in a discrete probability distribution.

Profile variation (PV) score

$$\mathrm{JSD}(P_1, P_2, \ldots, P_n) = H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i)$$

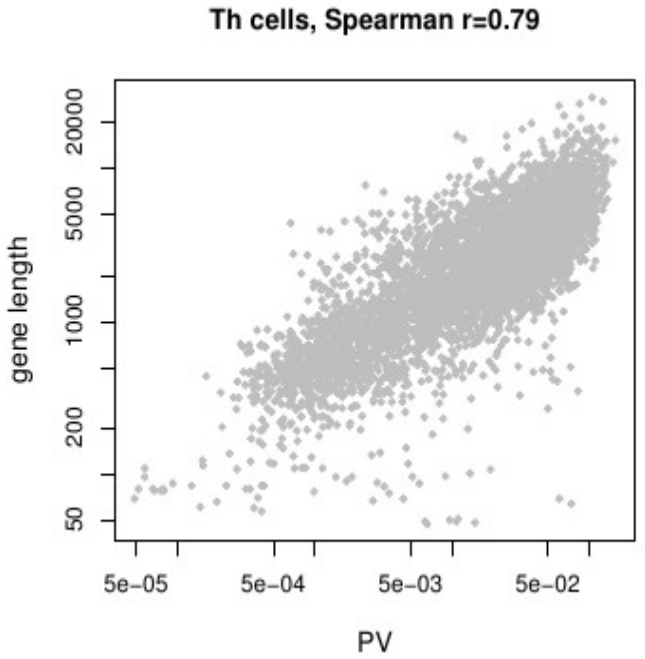
H( ): entropy

increases with the number of categories in a discrete probability distribution.

gene length

$$PV = JSD/\log_2(L)$$

PV with gene length regressed out



$$Y_{PV} = \beta_0 + \beta_1 X_{length}$$

$$\hat{Y_{PV}} = \hat{\beta_0} + \hat{\beta_1} X_{length}$$

$Y_{PV} - \hat{Y_{PV}}$ is the length regressed PV scores $PV_{\setminus length}$.

Outline

- Method
  - Profile Variation (PV) score --- two versions of PV score
- Benchmarking and thresholding
  - Various data sets
  - Various gene categories and exons
  - Compare with bulk RNA-seq
- Applications
  - Genes with high isoform variation
  - Patterns in isoform usage
  - Genes which switch isoforms

Compare between data sets

Differentiating T helper cells

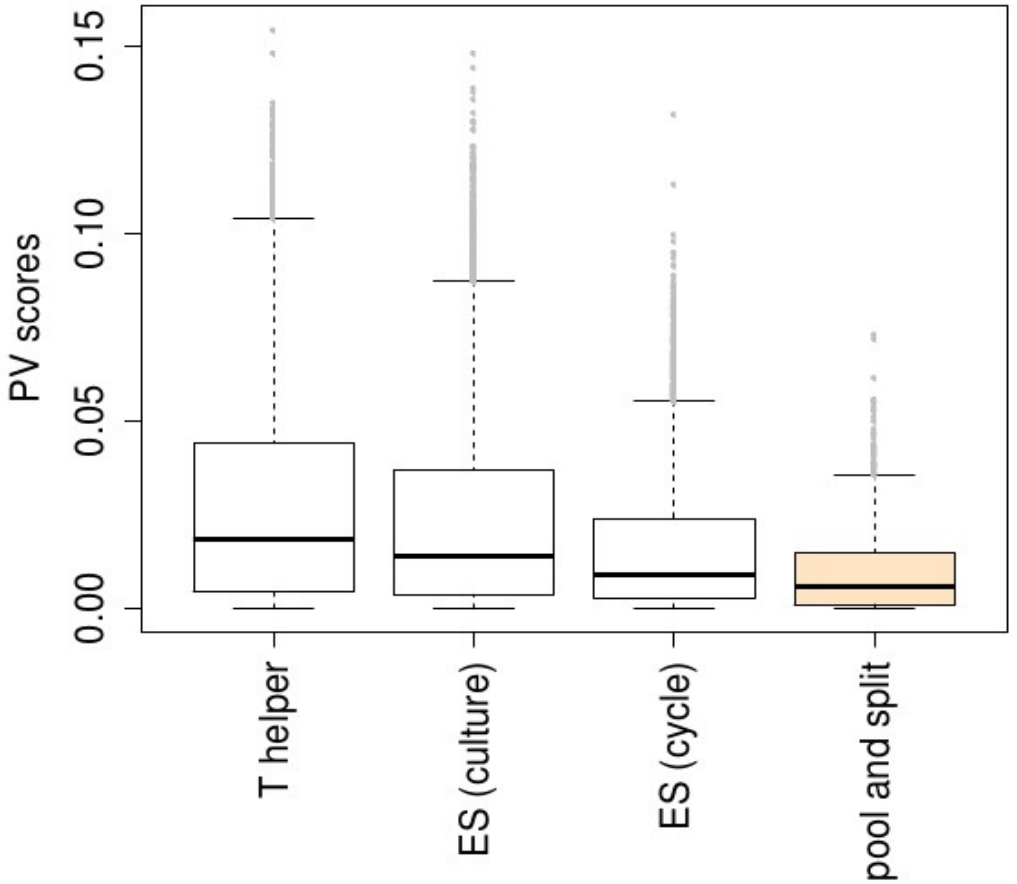
Mahata *et al* 2014

Embryonic stem (ES) cells: different culture conditions

Kołodziejczyk *et al* 2015

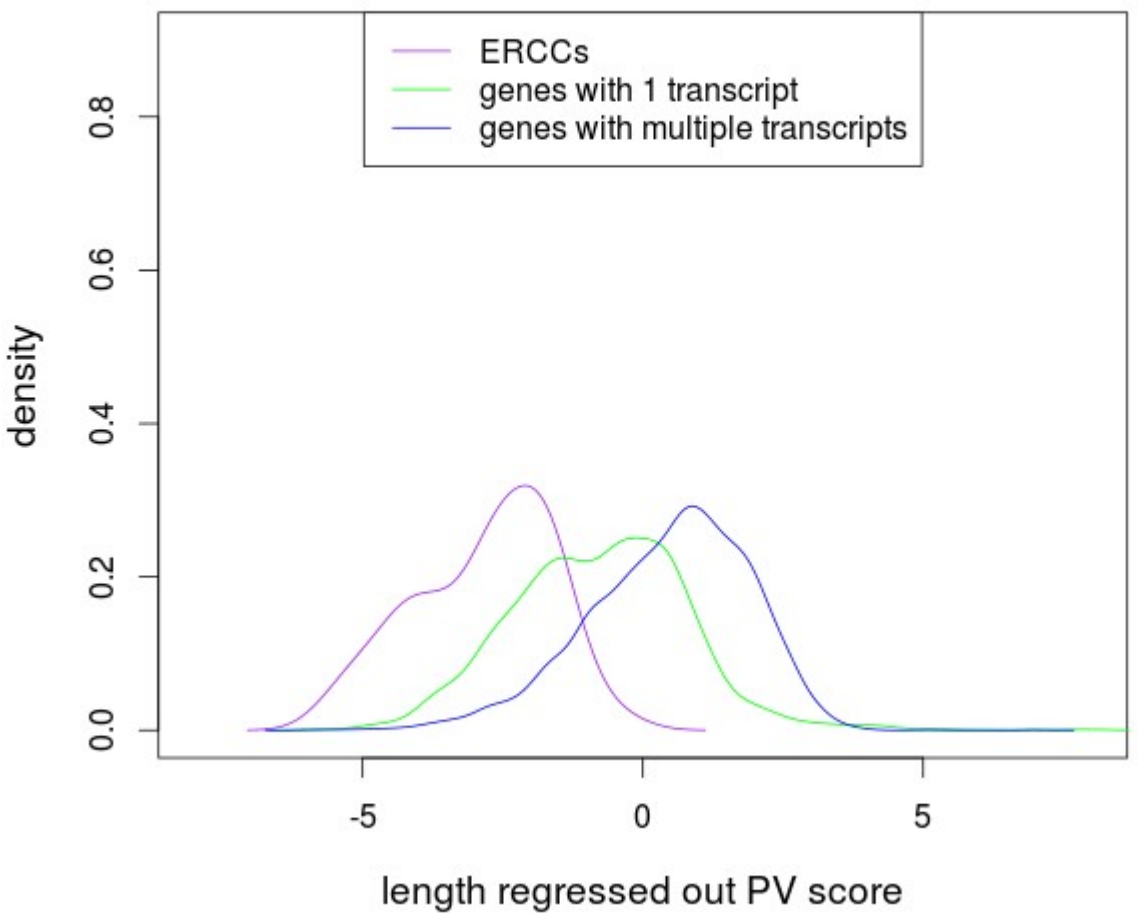
different cell cycle phases

Buttener *et al* 2015



Marinov *et al* 2014

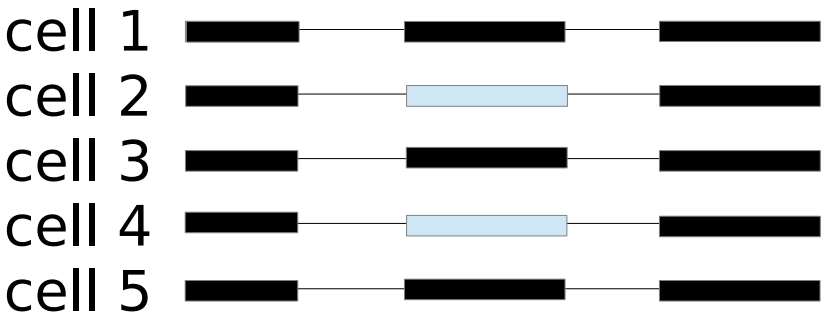
Compare between gene categories



Thresholding PV scores

PV of AS genes

- technical noise
- biological noise
- AS events



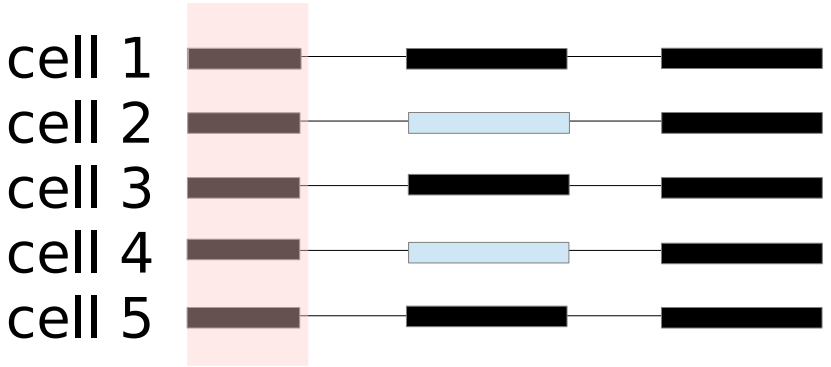
Thresholding PV scores

PV of AS genes

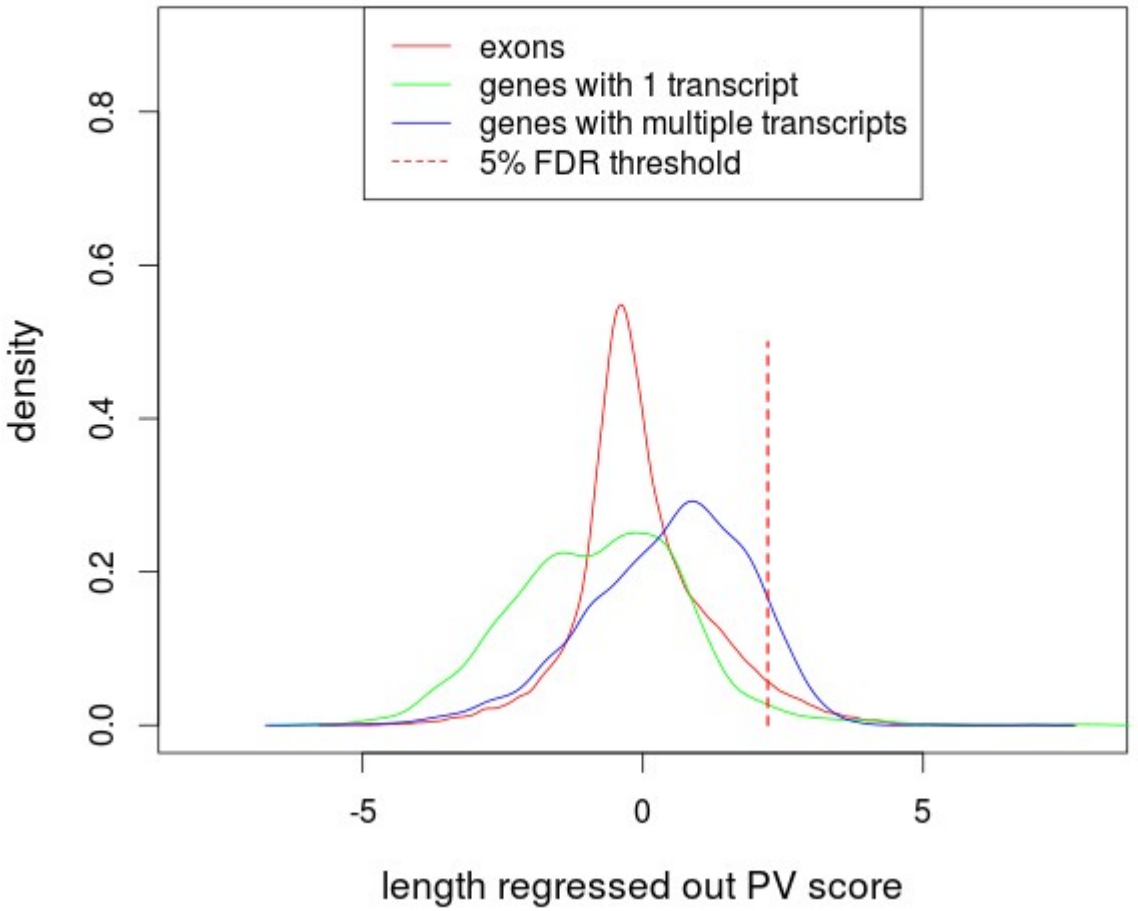
- technical noise
- biological noise
- AS events

PV of exons

- technical noise
- biological noise

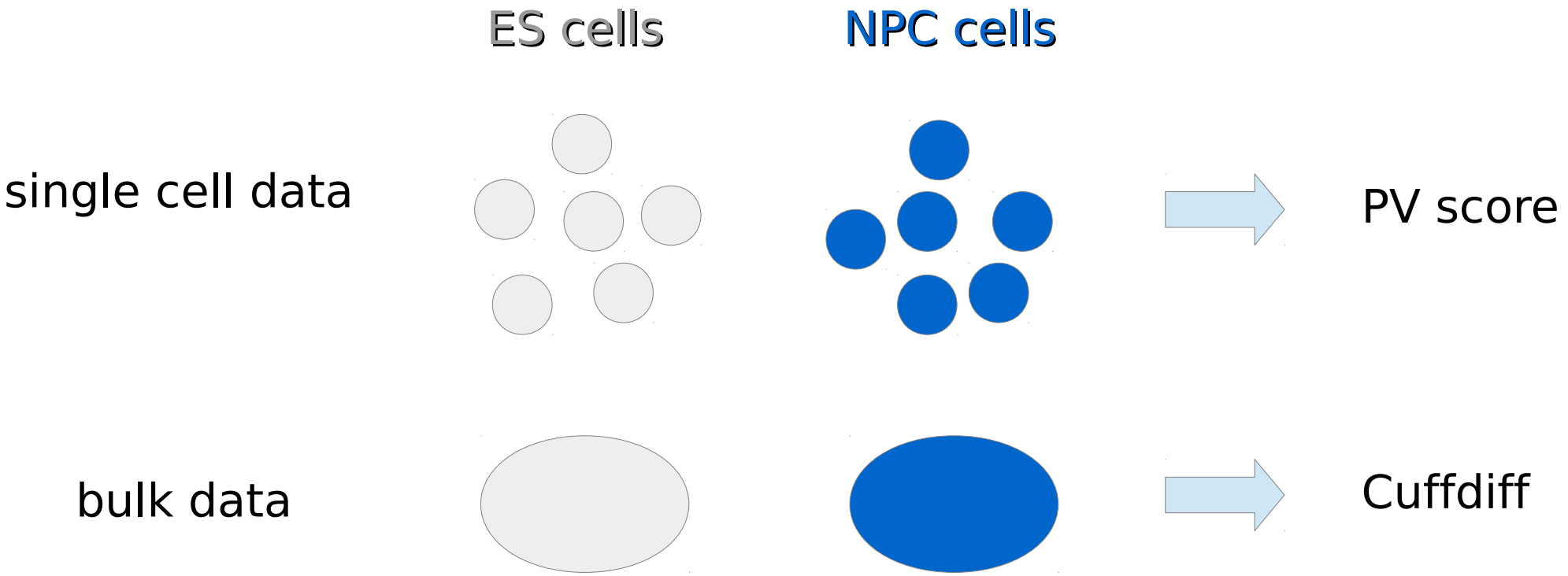


Thresholding PV scores with exons



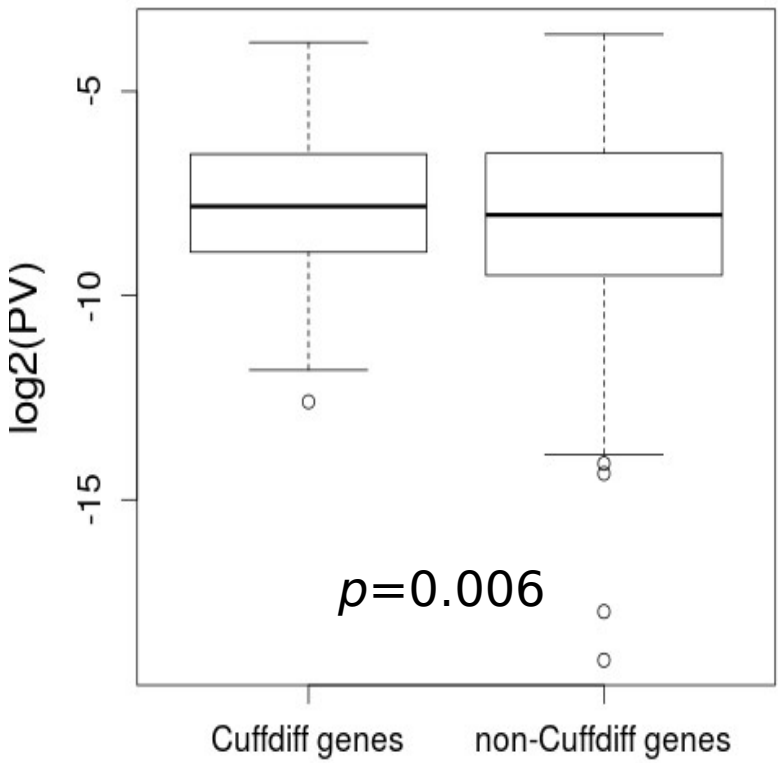
300~600 genes were found to have highly variable isoform usage

Compare with bulk RNA-seq data



Compare with bulk RNA-seq data

What is consistent



What is different

- Genes with high PV but not detected by Cuffdiff
- Enriched in cell cycle genes
- Biological variation within one cell type

Outline

- Method
  - Profile Variation (PV) score -- two versions of PV score
- Benchmarking and thresholding
  - Various data sets -- conforms with biological heterogeneity
  - Various gene categories and exons -- significant variation
  - Compare with bulk RNA-seq -- consistent and more than bulk
- Applications
  - Genes with high isoform variation
  - Patterns in isoform usage
  - Genes which switch isoforms

Genes with highly variable isoforms

Isoform variation at two levels

PV

expression regulation
chromatin modification

$PV_{\backslash length}$

immunology

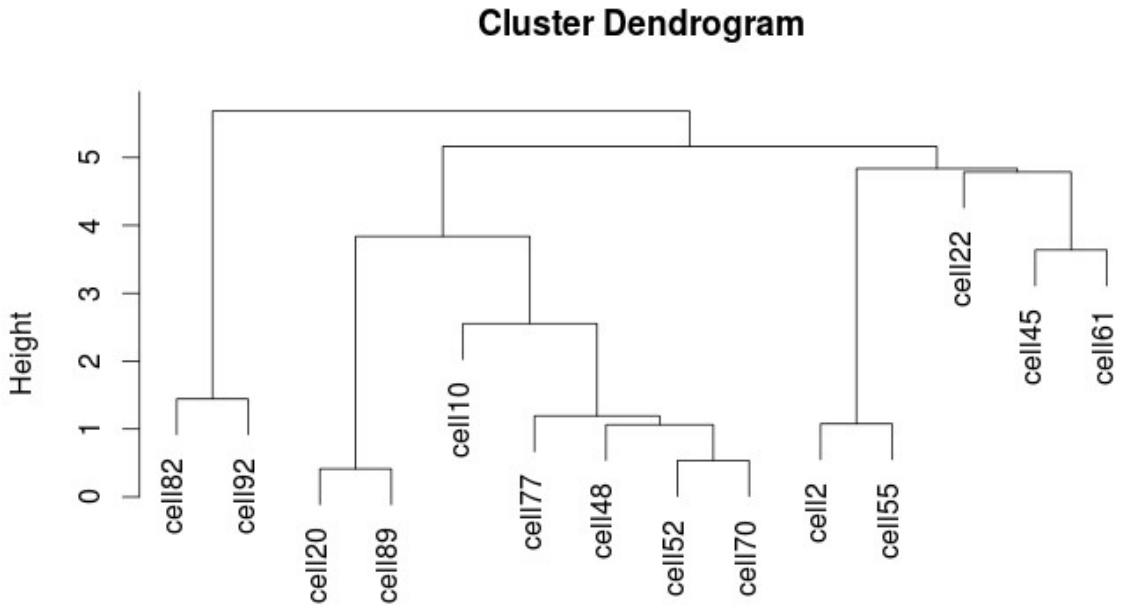
T helper cells

cell cycle

ES cells

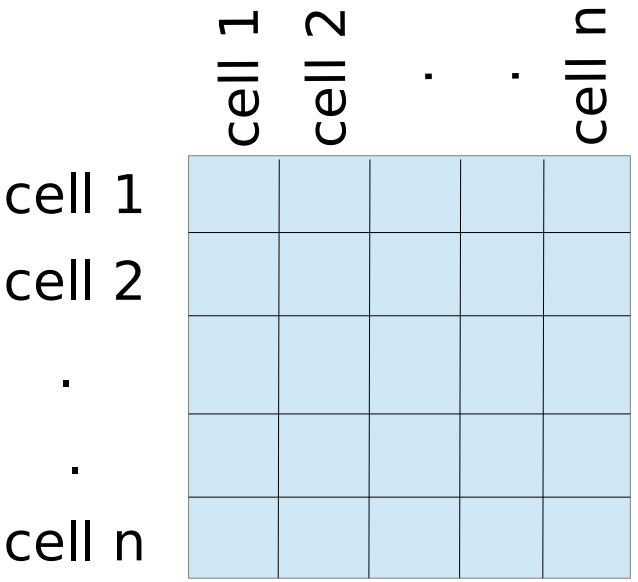
Find representative read coverage patterns

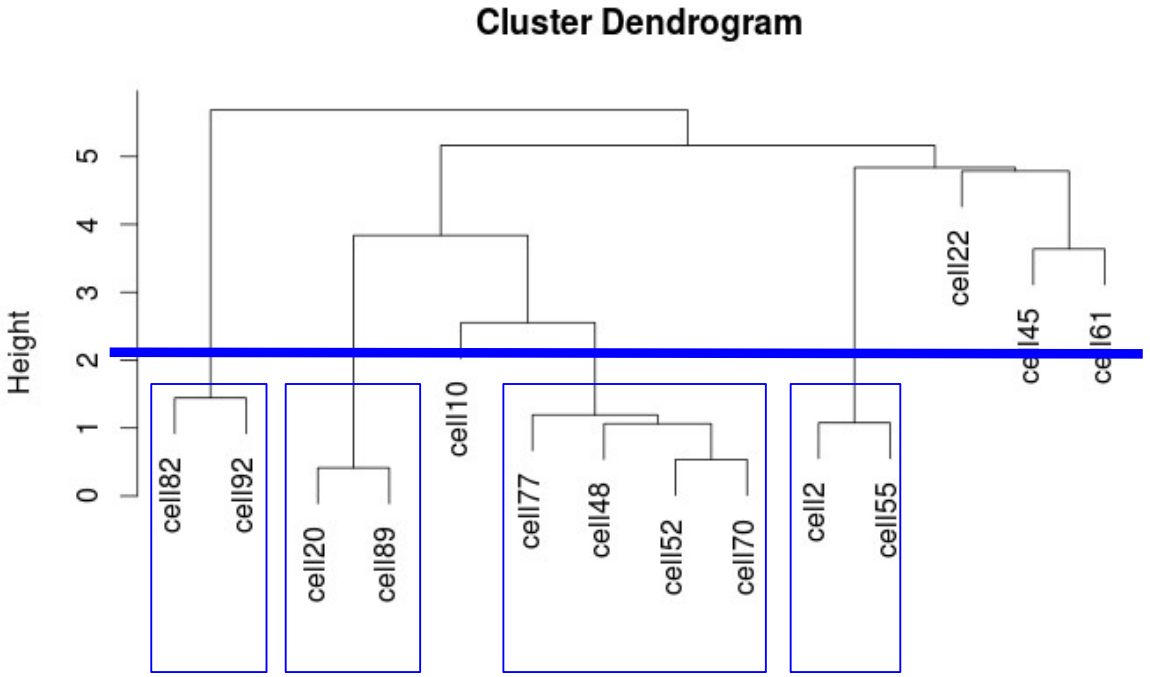
	cell 1	cell 2	.	.	cell n
cell 1		$\sqrt{PV}$			
cell 2					
.					
.					
cell n					



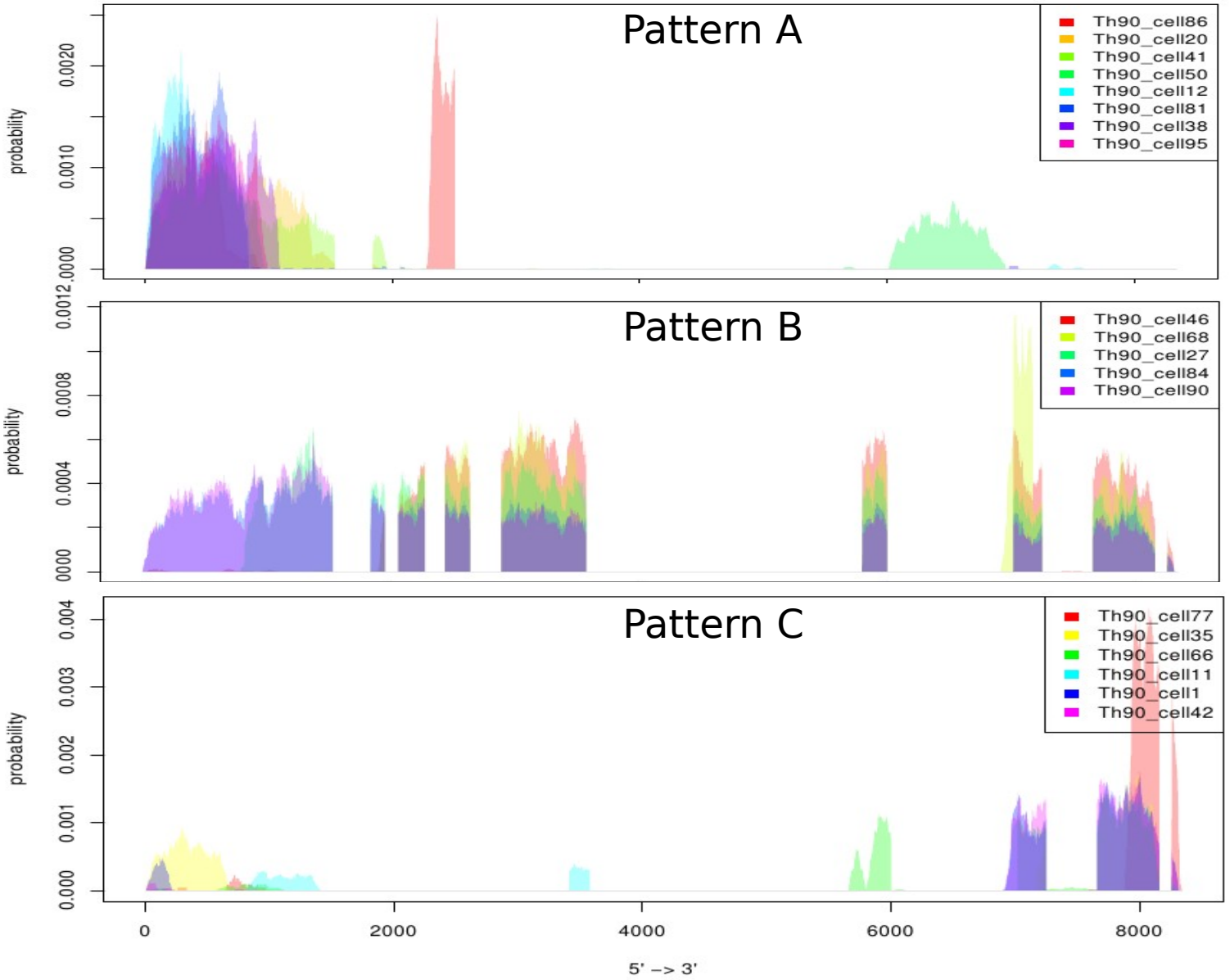
pairwise $\sqrt{PV}$ is a metric

Find representative read coverage patterns





Find representative read coverage patterns



Example: *Nsf* in T cells

Find correlated genes in isoform usage

	Gene 1	Gene 2	Gene 3
Cell 1	Pattern A	Pattern A	Pattern A
Cell 2	Pattern A		Pattern A
Cell 3			Pattern C
Cell 4	Pattern B	Pattern B	Pattern B
Cell 5	Pattern B	Pattern B	

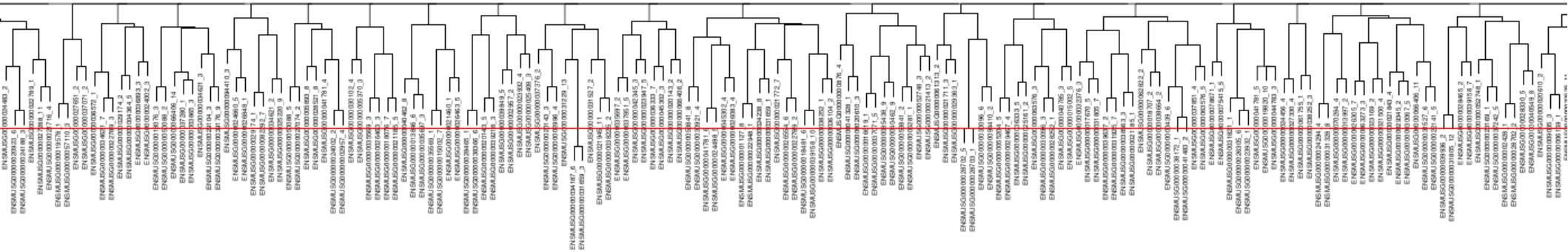
Difficulty: genes are expressed in a small number of cells.

Find correlated genes in isoform usage

Cell *vs* “gene pattern“ binary matrix

	Gene1 PatternA	Gene1 PatternB	Gene2 PatternA	Gene2 PatternB	Gene3 PatternA	Gene3 PatternB	Gene3 PatternC
Cell 1	1	0	1	0	1	0	0
Cell 2	1	0	0	0	1	0	0
Cell 3	0	0	0	0	0	0	1
Cell 4	0	1	0	1	0	1	0
Cell 5	0	1	0	1	0	0	0

Stochasticity in isoform usage

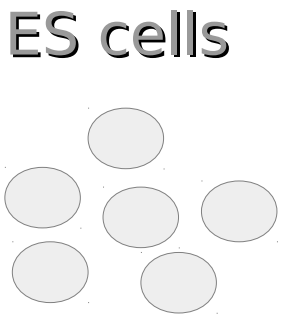


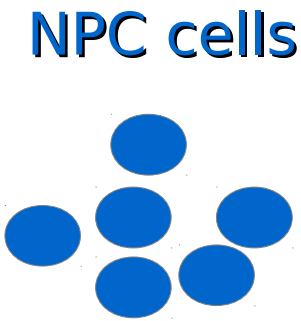
Find clusters of genes with *Jaccard distance* < *h*

Compare with random binary matrices:

Isoform usage across single cells has high stochasticity

Genes which switch isoforms between cell types





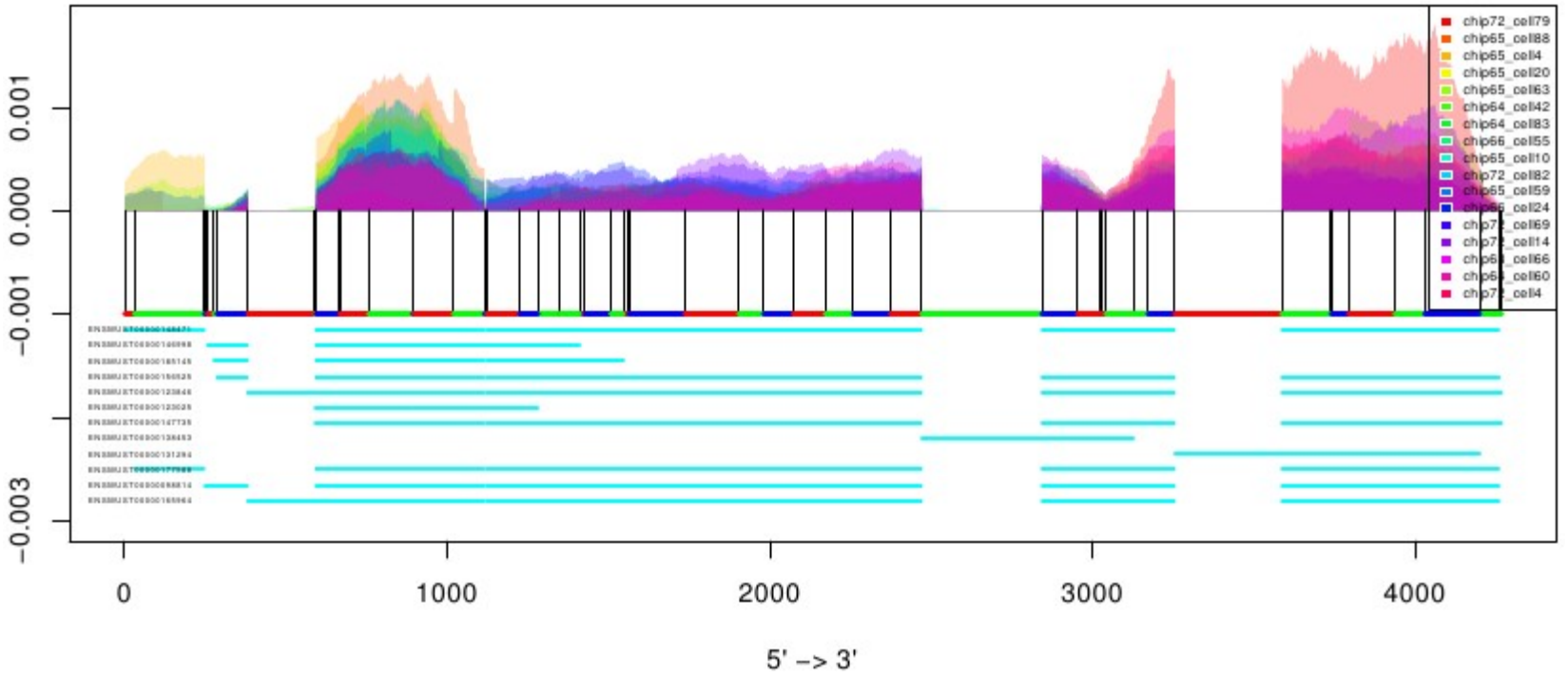
	ES-cell1	ES-cell2	...	...	ES-cellm	NPC-cell1	NPC-cell2	...	NPC-celln
ES-cell1	Intra group distances					Inter group distances			
ES-cell2									
...									
ES-cellm									
NPC-cell1	Inter group distances					Intra group distances			
NPC-cell2									
...									
NPC-celln									

1. high ratio of:

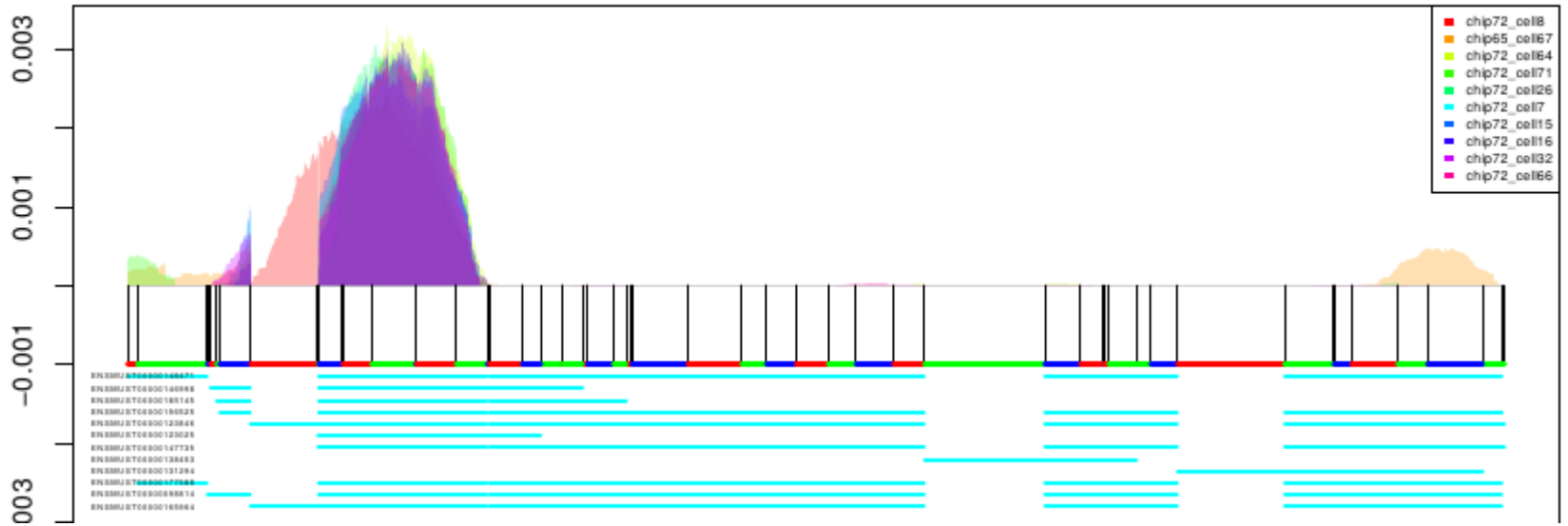
$$\frac{\text{Average(Inter-group distances)}}{\text{Average(Intra-group distances)}}$$

2. high PV(all cells)

Lig1

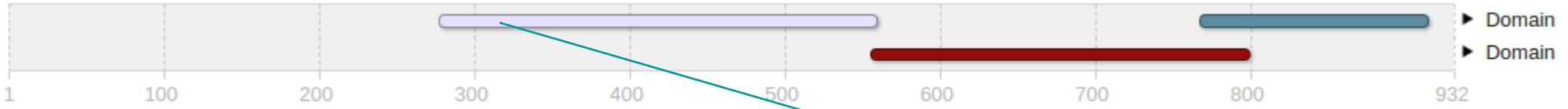


12 ES cells
5 NPC cells

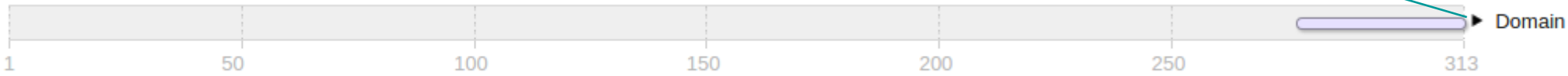


NPC cells

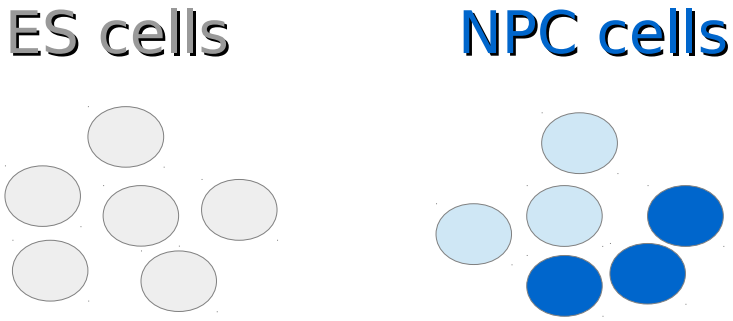
Lig1
protein domains of the long transcript



Lig1
protein domain of the short transcript



- Lig1 is a cell cycle gene
- Cells slow down cycling ES → NPC
- Not detected in bulk data



Outline

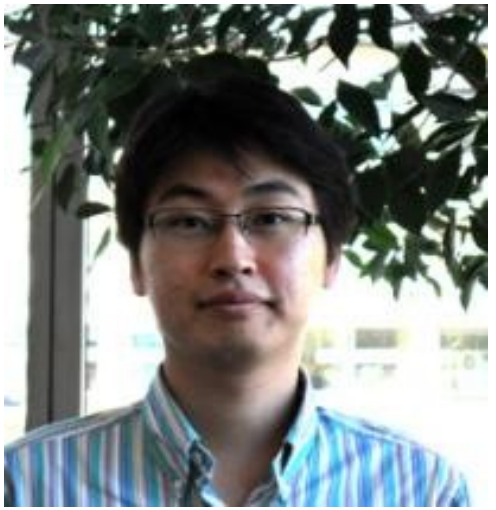
- Method
  - Profile Variation (PV) score -- two versions of PV score
- Benchmarking and thresholding
  - Various data sets -- conforms with biological heterogeneity
  - Various gene categories and exons -- significant variation
  - Compare with bulk RNA-seq -- consistent and more than bulk
- Applications
  - Genes with high isoform variation -- sources of isoform variation
  - Patterns in isoform usage -- high stochasticity
  - Genes which switch isoforms -- function change during differentiation

Teichmann Group

Acknowledgements

Sarah Teichmann

Valentine Svensson



Jong Kyoung Kim

EMBL-EBI

Oliver Stegle (EBI)

John Marioni (EBI)

Nick Owens (MRC NIMR)

Kaur Alasoo (Sanger Institute)

FNSNF

Swiss National Science Foundation