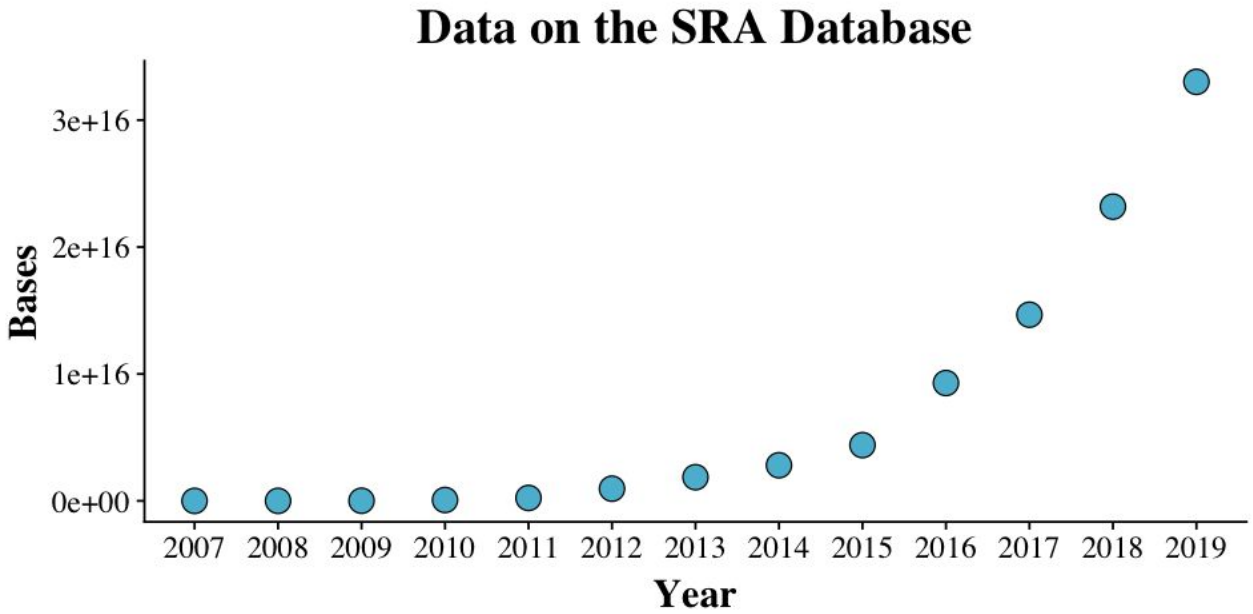
Rocky2019 ISCB Conference

Discrete Wavelet Transforms as a Batch Correction Tool

Rutendo F. Sigauke
Dowell Lab
rutendo.sigauke@colorado.edu

December 6, 2019

There is an exponential growth of sequencing data



SRA is a great resource for meta-analysis studies

Leinonen et al. Nucleic Acids Research 2010

Background and Motivation

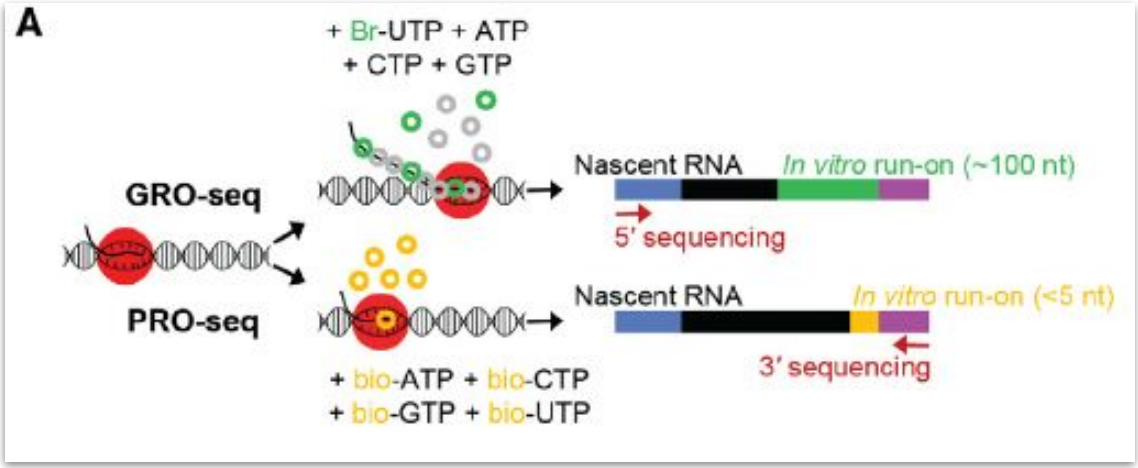
Single lab - single experiment

Sample	Condition
WT_replicate_1	Control
WT_replicate_2	Control
Treatment_replicate_1	Treatment
Treatment_replicate_2	Treatment

Single lab - multiple experiments

Sample	Condition	Batch
WT_replicate_1	Control	1
WT_replicate_2	Control	2
Treatment_replicate_1	Treatment	1
Treatment_replicate_2	Treatment	2

There are two main protocols for nascent RNA-seq

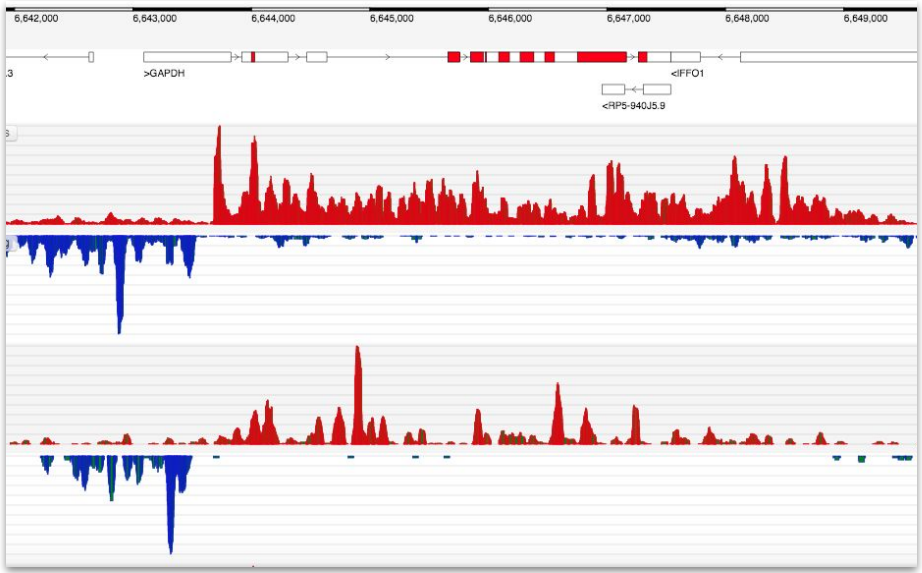


Random priming **Circularization** **Ligation**

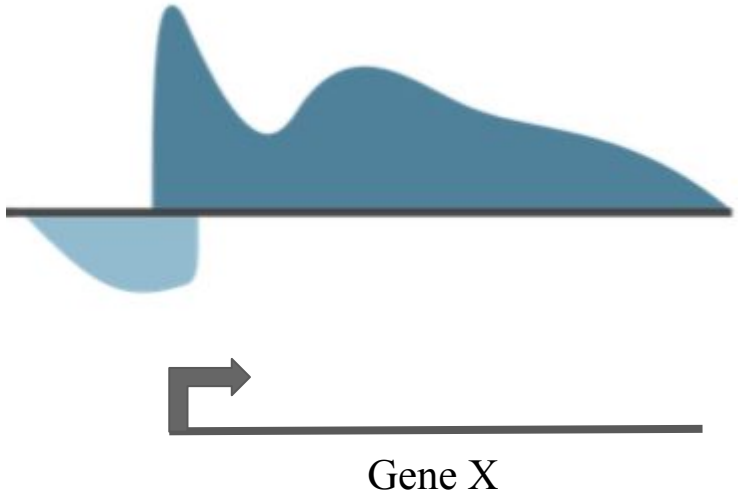
Wang et al. Cell Reports 2014

Real coverage data compared to simplified model

GAPDH GRO-seq Coverage



Theoretical model of GRO-seq coverage

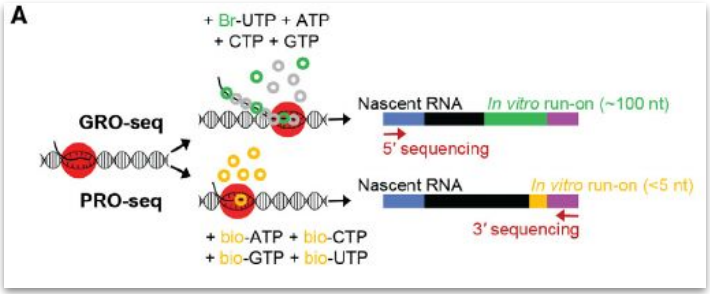


https://nascent.colorado.edu/browser/

David Deen:www.daviddeen.com

Example of noise in datasets we can explore

Samples	Projects/Labs	Protocols	Cell lines
102	17	2	8



HCT116

MCF7

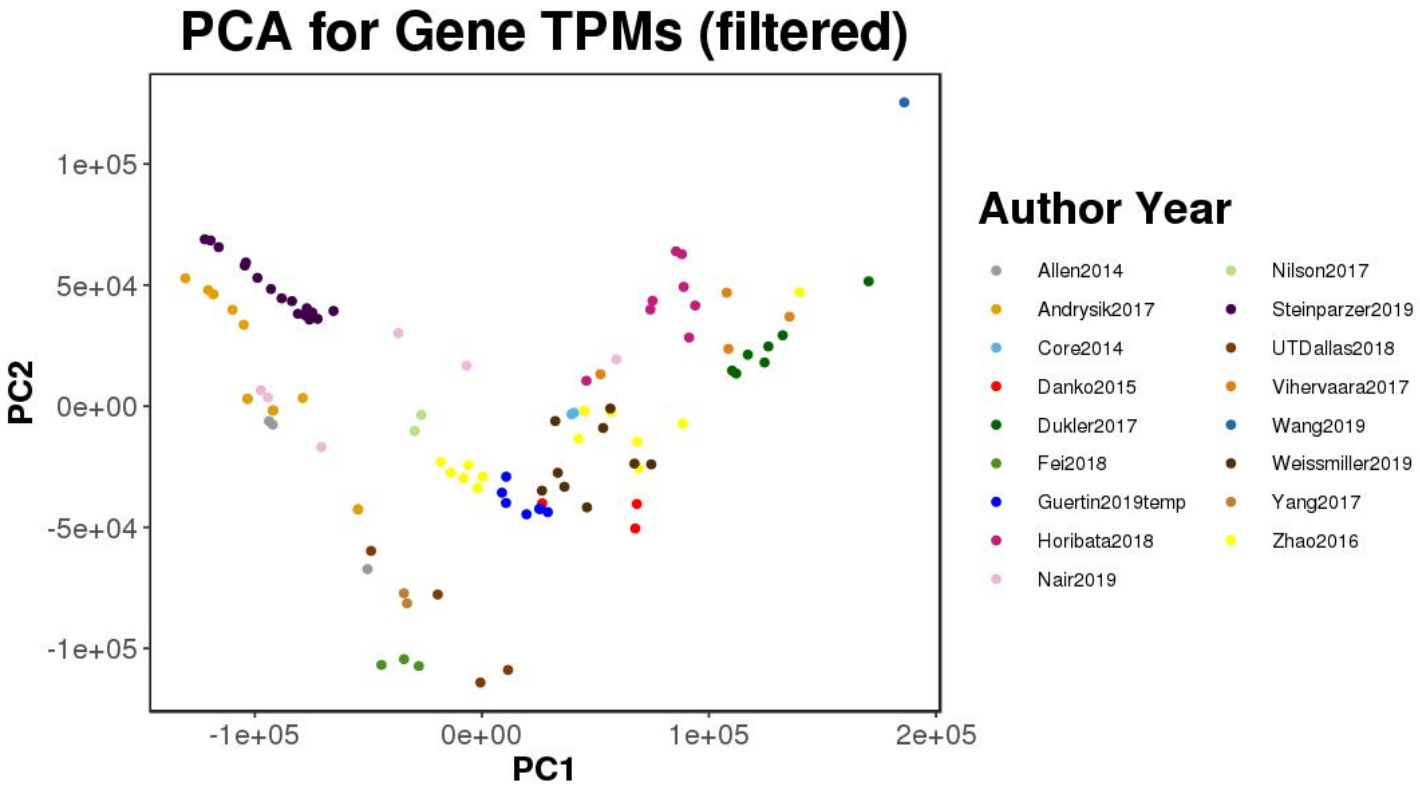
K562

SJSA

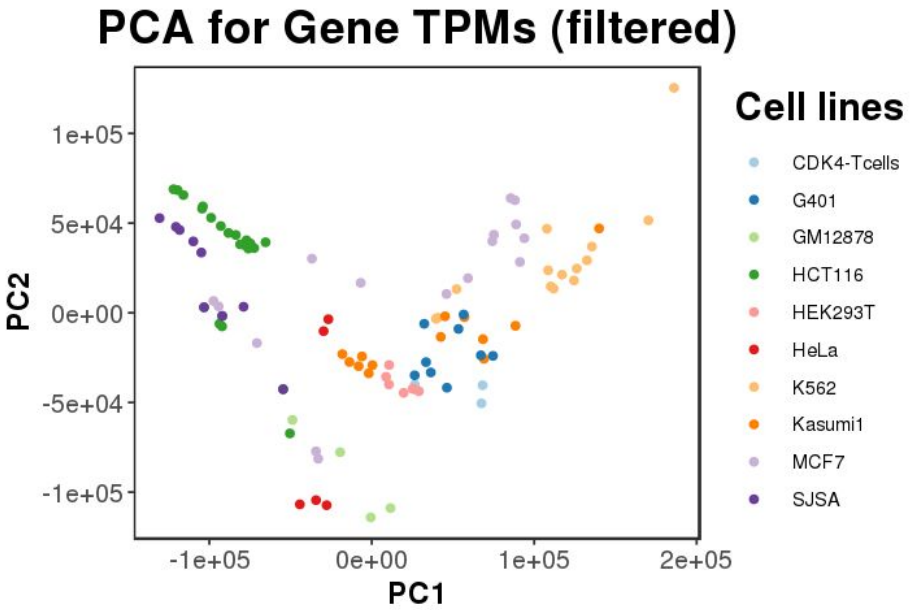
AC16

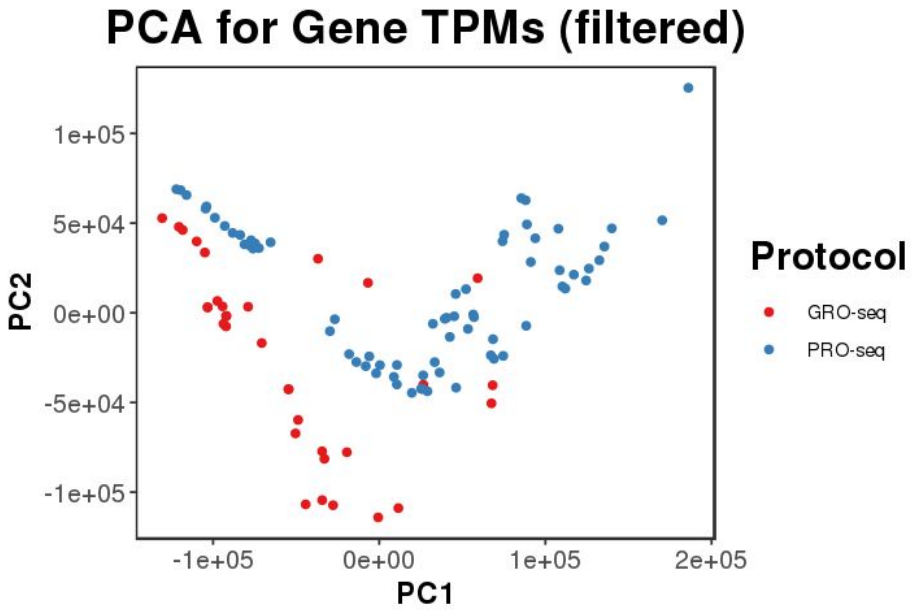
GM12878

There are batch effects specific to lab and protocol



There are batch effects specific to lab and protocol





The biological question of interest is cell line specific

...what if we can remove **technical noise** specific to batch, protocol etc. using wavelets...?

Why use wavelets?

1. Does not require a balanced experimental design
2. The actual data is not altered
3. We do not have to specify “confounders” ie. batch, protocol

Mathematical Representation of a Wavelet Transform

$$X_{g(a,b)} = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} g(t)\psi \frac{t-b}{a} dt$$

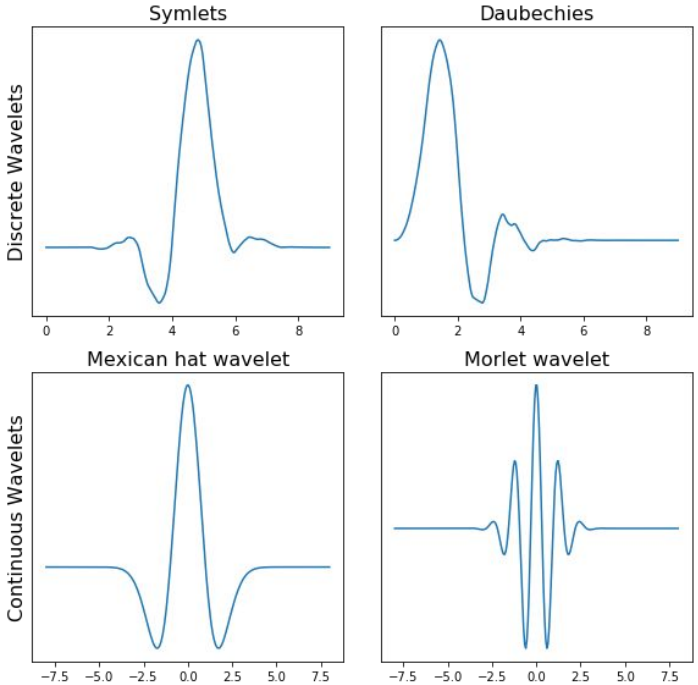
$\psi(t)$: mother wavelet

a : scaling factor

b : translation factor

DWT uses discrete values for the scale (a) and translation factor (b)

- the scale factor increases in powers of two ($a = 1, 2, 4. . .$)
- the translation factor increases integer values ($b = 1, 2, 3. .$)

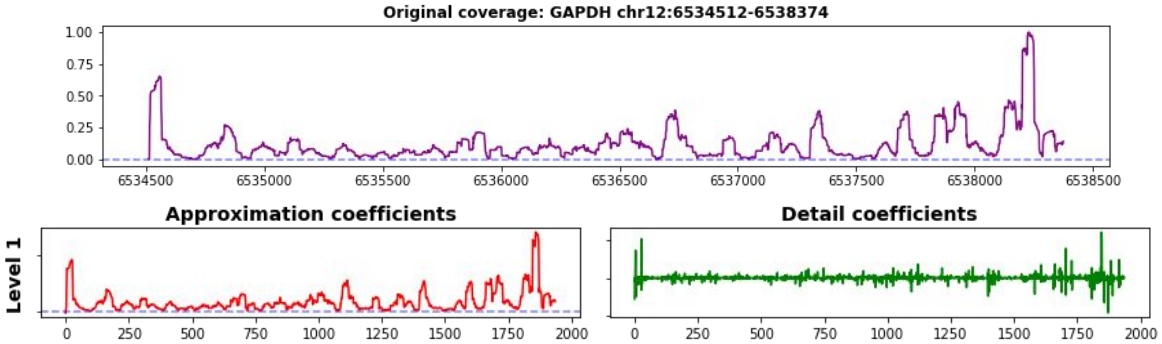


Current workflow for denoising nascent RNA-seq

1. Select genes

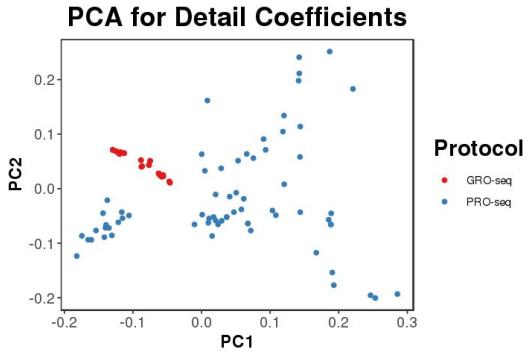
Housekeeping genes

2. Normalize transcription levels



3. Discrete wavelet transformation of coverage

4. Cluster on the detail coefficients



Limitations

1. Library depth affects the sampling of lowly transcribed transcripts
2. Low complex libraries affect the significantly detail coefficients
3. Alternative transcription/expression of genes with also affect the transforms

Proposed Solutions

1. *Downsample all libraries to a minimum depth*
2. *Filter libraries with low complexity and depth*
3. *Select genes without multiple isoforms*

Acknowledgements

Dowell and Allen Lab

Robin Dowell, D.Sc.
Mary Allen, Ph.D.
Jacob Stanley, Ph.D.
Margaret Gruca, M.A.

BioFrontiers IT