Using statistical properties of gene expression data to improve inference and classification

Aedan Roberts

PhD candidate
School of Software & Centre for Artificial Intelligence
Faculty of Engineering and IT
University of Technology Sydney

SoS HDR Showcase, 2nd April 2019

Outline

What is gene expression data

Gene expression

- Measure of "activity" of a gene
- Activity of genes controls behaviour of cells, tissues, organs
- Abnormal gene expression $\rightarrow$ loss of normal function

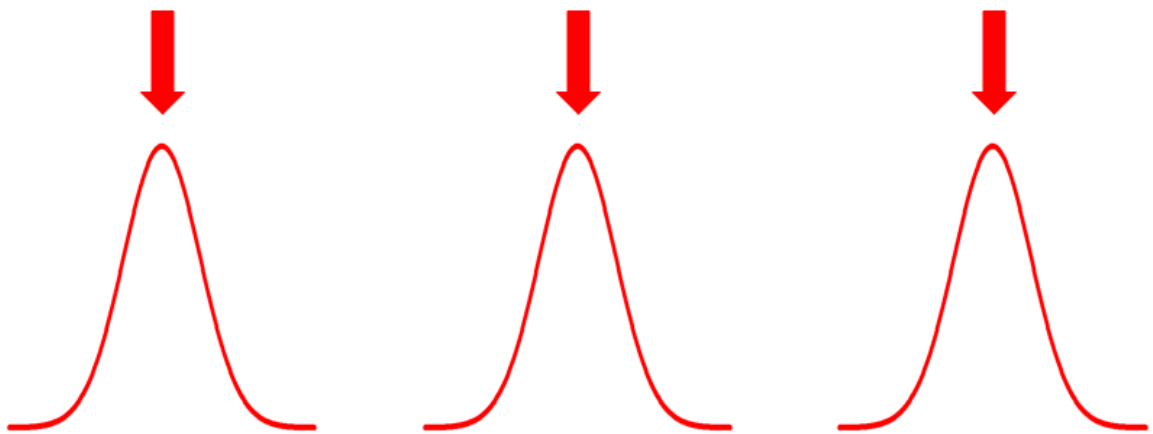
Gene expression data

- Activity of all genes in a sample at a given time
- High dimensional: $\sim$ 30,000 genes per sample

Gene expression data analysis

Expression values for each patient → distribution for each gene

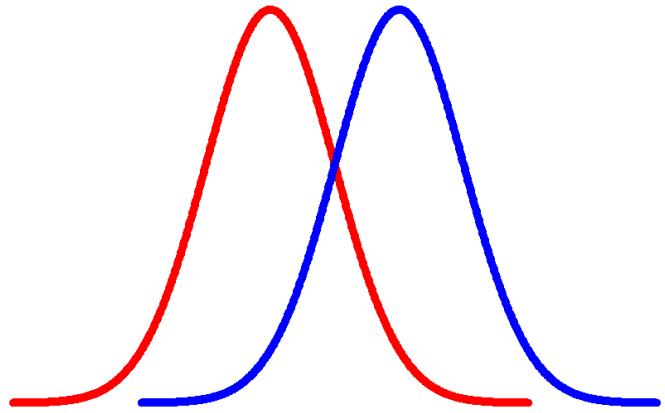
	Gene 1	Gene 2	Gene 3	...
Patient 1	10	4	33	...
Patient 2	15	7	45	...
Patient 3	18	6	15	...
Patient 4	15	5	15	...
Patient 5	22	1	66	...
...	...	...	...	...



Gene expression analysis

For each gene, compare distributions between groups, e.g.

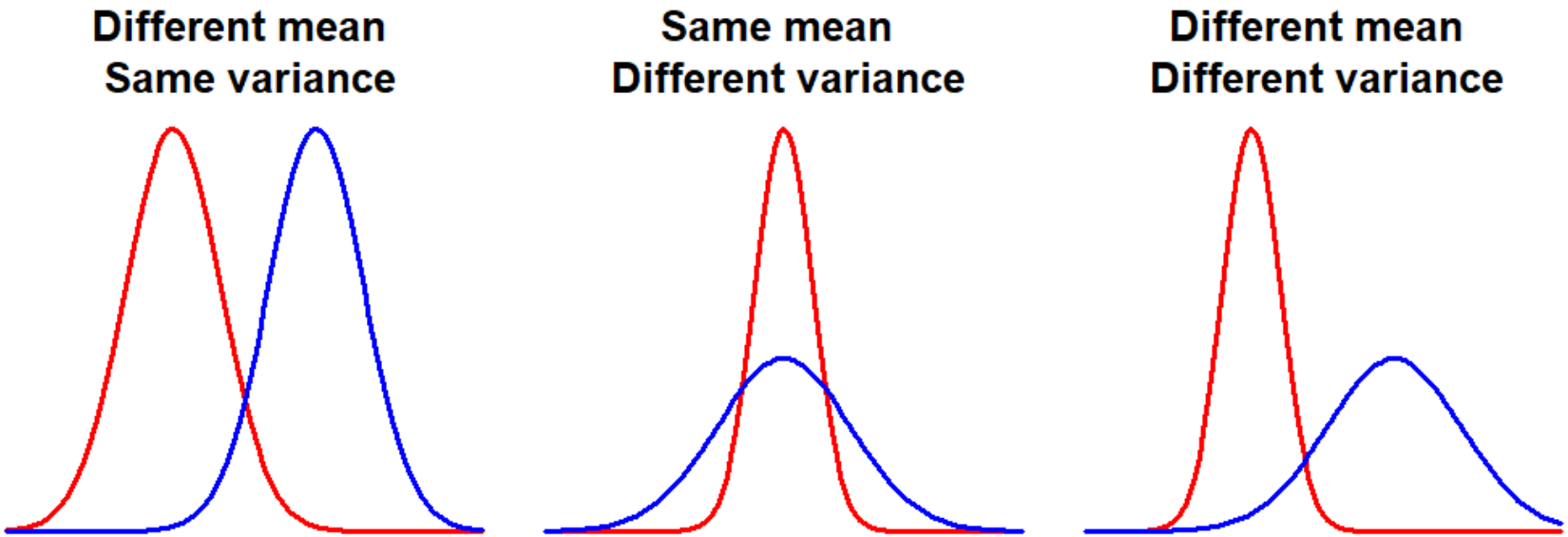
- Cancer v normal
- Different subtypes of cancer



Aim: Identify genes with different distributions in each group

- For biological interest – understand disease, develop treatments
- For classification – improve diagnosis or treatment decisions

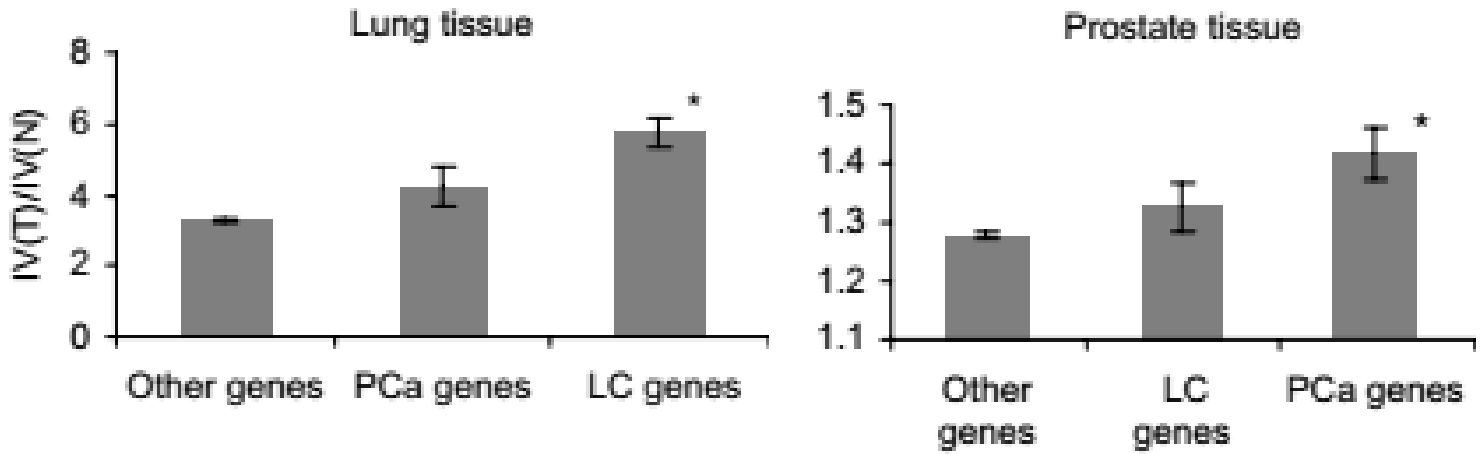
Limitations of existing methods



Most methods only look for differences in mean, but there is evidence that differences in variance are also relevant

Changes in variance are biologically relevant

$$\textit{Variance ratio}: \quad \frac{\text{Variance among cancer samples}}{\text{Variance among normal samples}}$$



Gorlov *et al.* (2012) J Bioinf Comp Biol

- Variance higher in tumour compared to normal tissue
- Increase is greater for genes related to specific cancer type

Cancer classification using differences in variance

Previous work shows that genes with increased variance can be used for classification

- Limited to cancer *v* normal or more *v* less aggressive
- Increased variance in cancer or in more aggressive form

Can differential variance be applied more widely?

- Classification of cancer subtypes where there is no expected general increase in variance in one group

Cancer classification using differences in variance

Test on publicly available cancer gene expression data

- Three classification tasks on two pairs of datasets
- Select features by differences in mean, variance or both
- Compare classification performance

Dataset	-ve/+ve classes	-ve samples	+ve samples
GSE13351	Low/high hyperdiploid	64	28
GSE79533	Low/high hyperdiploid	136	90
GSE19475	MLL-AF4/MLL-ENL	28	22
GSE68720	MLL-AF4/MLL-ENL	48	16
GSE19475	MLL translocation -/+	14	58
GSE68720	MLL translocation -/+	17	80

Cancer classification using differences in variance

Comparison	Differential variance			Differential expression			Combined feature selection		
	RF	FLDA	SVM	RF	FLDA	SVM	RF	FLDA	SVM
Hyperdiploidy	0.963	0.967	*0.975*	0.970	0.966	***0.984***	*0.971*	0.969	0.970
	0.920	*0.953*	0.925	*0.978*	0.977	0.967	0.981	***0.983***	0.980
MLL type	0.939	0.947	***0.958***	0.922	0.924	*0.956*	0.935	0.939	*0.949*
	0.988	0.987	***0.997***	0.855	*0.943*	0.924	0.948	0.959	*0.984*
MLL +/-	0.908	*0.921*	0.873	0.938	0.939	***0.949***	0.938	*0.942*	0.921
	0.980	0.954	0.955	***0.999***	0.998	0.978	0.997	*0.998*	0.985

AUC. RF = random forest; FLDA = Fisher's linear discriminant analysis; SVM = support vector machine (linear for differential expression, RBF kernel for others)

Conclusions

- Features selected by differences in variance can be used to classify cancer subtypes
- "Differential distribution" often better than differences in mean or variance alone

Limitations of existing work

Not applicable to small samples

- Need to be able to accurately estimate parameters
- Often have small sample sizes – maybe 5 or 10 samples
- Standard methods don't work well for small samples

Not applicable to modern gene expression data

- Previous work on microarray data – expression values approximately normally distributed
- **RNA-seq**: newer technology, more accurate, but can't assume normal distribution

→ Aim: accurately estimate parameters from small-sample RNA-seq data

Hierarchical models for inference in small samples

- Assume parameters for each gene are related to parameters for other genes
- Allows information from all genes to contribute to estimates for each individual gene

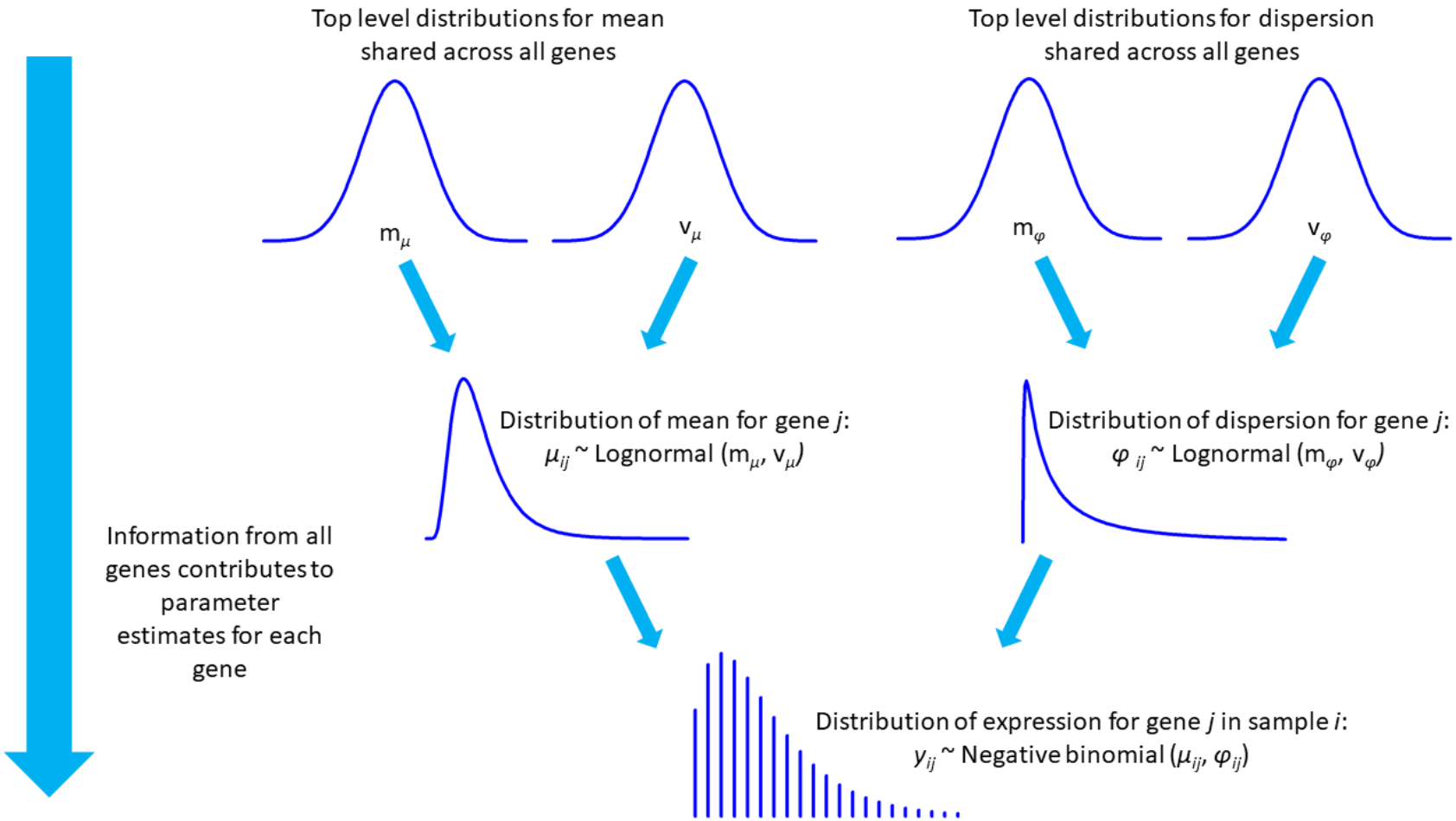
Expression value for gene j in sample i follows Negative Binomial distribution with mean μ_j and dispersion ϕ_j:

$$y_{ij} \sim NB(\mu_j, \phi_j)$$

Means and dispersions for each gene related by a common higher-level Lognormal distribution:

$$\mu_j \sim LN(m_\mu, v_\mu),\ \phi_j \sim LN(m_\phi, v_\phi)$$

Hierarchical model for RNA-seq data



Inference from hierarchical model

Parameters estimated by Bayesian inference

- Priors for mean and dispersion estimated from published data
- Priors combined with data to estimate posterior distributions for each mean and dispersion
- Estimates of mean and dispersion for each gene from posteriors

Posterior distributions for mean and dispersion are not standard probability distributions

- Can't make direct inferences from posteriors
- $\rightarrow$ Sample using Markov chain Monte Carlo

Progress to date

- Hierarchical model for parameter estimation built and optimised
- Tested against existing methods for mean and dispersion estimation using data simulated under different models:

Data	Model	MSE μ	MSE ϕ
Lognormal	Hierarchical model	30.32	0.0651
	edgeR	38.11	0.0695
	DESeq2	36.35	0.0914
Gamma	Hierarchical model	15.61	0.1089
	edgeR	18.45	0.0648
	DESeq2	18.11	0.0746

Mean squared errors for mean and dispersion for simulated gene expression data for 10,000 genes in 10 samples.

Next steps

- Test hierarchical model on real data
- Extend model to identify genes with differences in distribution between groups
  - Mixture model – extra hierarchical layer to estimate probability that mean or dispersion is different for each group
- Apply features identified by mixture model to classification of cancer subtypes
- Apply to single-cell RNA-seq data

Acknowledgements

Supervisors

A/Prof Paul Kennedy
A/Prof Daniel Catchpoole

Funding

NSW Health PhD Scholarship