Many regressions simultaneously

Jeff Leek

@jtleek

www.jtleek.com

Genomics measures many variables

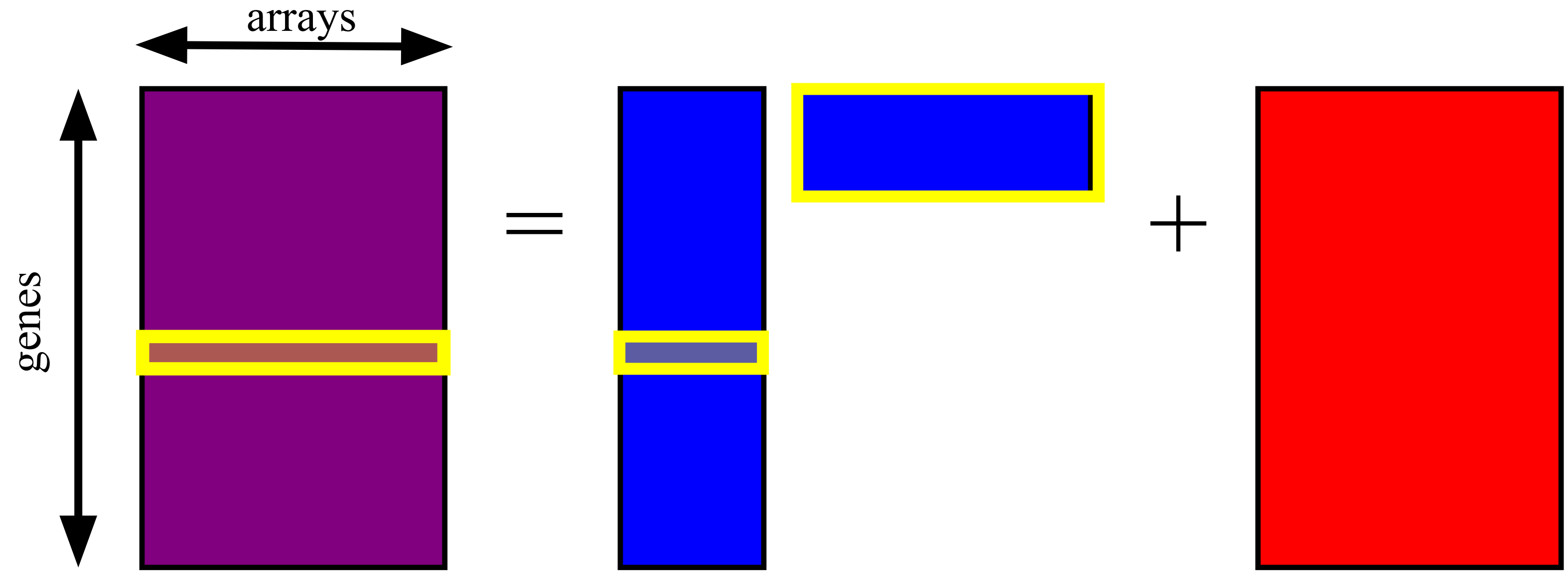
Slide adapted from: John Storey

VARYING CONDITIONS

	sample 1	*sample 2*	*sample 3*	…	*sample n*
feature 1	1.23	2.37	0.345	…	4.27
feature 2	3.98	1.09	0.237	…	0.283
⋮		⋮			⋮
feature m	0.896	2.81	3.92	…	1.46

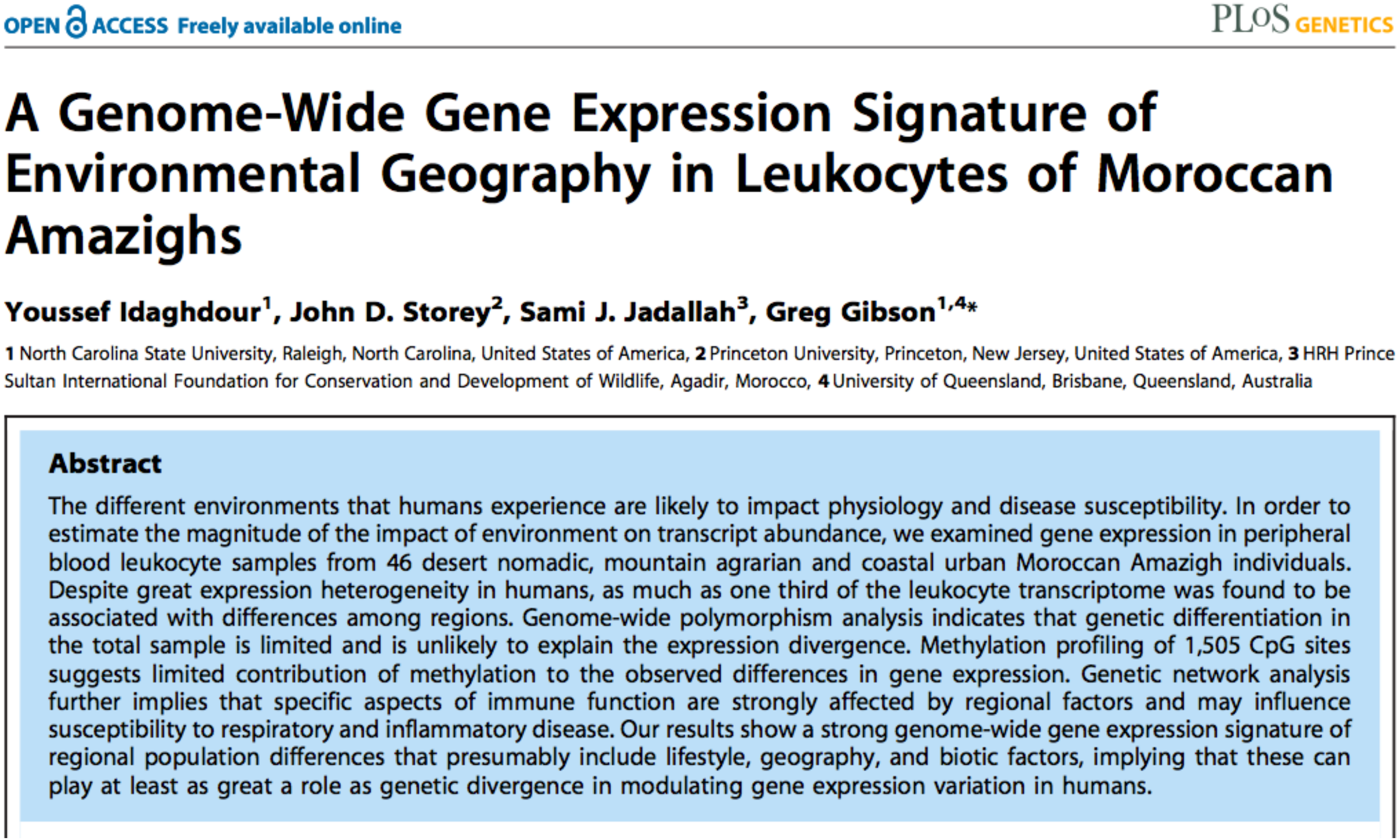
A common goal is to find which of them relate to an outcome

Slide adapted from: John Storey



$$\mathbf{X} = \mathbf{B} \quad \mathbf{S(Y)} + \mathbf{E}$$

Slide adapted from: John Storey



OPEN ACCESS Freely available online

PLoS GENETICS

A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs

Youssef Idaghdour[1], John D. Storey[2], Sami J. Jadallah[3], Greg Gibson[1,4]*

1 North Carolina State University, Raleigh, North Carolina, United States of America, **2** Princeton University, Princeton, New Jersey, United States of America, **3** HRH Prince Sultan International Foundation for Conservation and Development of Wildlife, Agadir, Morocco, **4** University of Queensland, Brisbane, Queensland, Australia

Abstract

The different environments that humans experience are likely to impact physiology and disease susceptibility. In order to estimate the magnitude of the impact of environment on transcript abundance, we examined gene expression in peripheral blood leukocyte samples from 46 desert nomadic, mountain agrarian and coastal urban Moroccan Amazigh individuals. Despite great expression heterogeneity in humans, as much as one third of the leukocyte transcriptome was found to be associated with differences among regions. Genome-wide polymorphism analysis indicates that genetic differentiation in the total sample is limited and is unlikely to explain the expression divergence. Methylation profiling of 1,505 CpG sites suggests limited contribution of methylation to the observed differences in gene expression. Genetic network analysis further implies that specific aspects of immune function are strongly affected by regional factors and may influence susceptibility to respiratory and inflammatory disease. Our results show a strong genome-wide gene expression signature of regional population differences that presumably include lifestyle, geography, and biotic factors, implying that these can play at least as great a role as genetic divergence in modulating gene expression variation in humans.

Slide adapted from: John Storey

Primary (“Biological”) Variable

A	**A**	**D**	**V**	**…**	**D**	**A**	**V**

Adjustment Variables

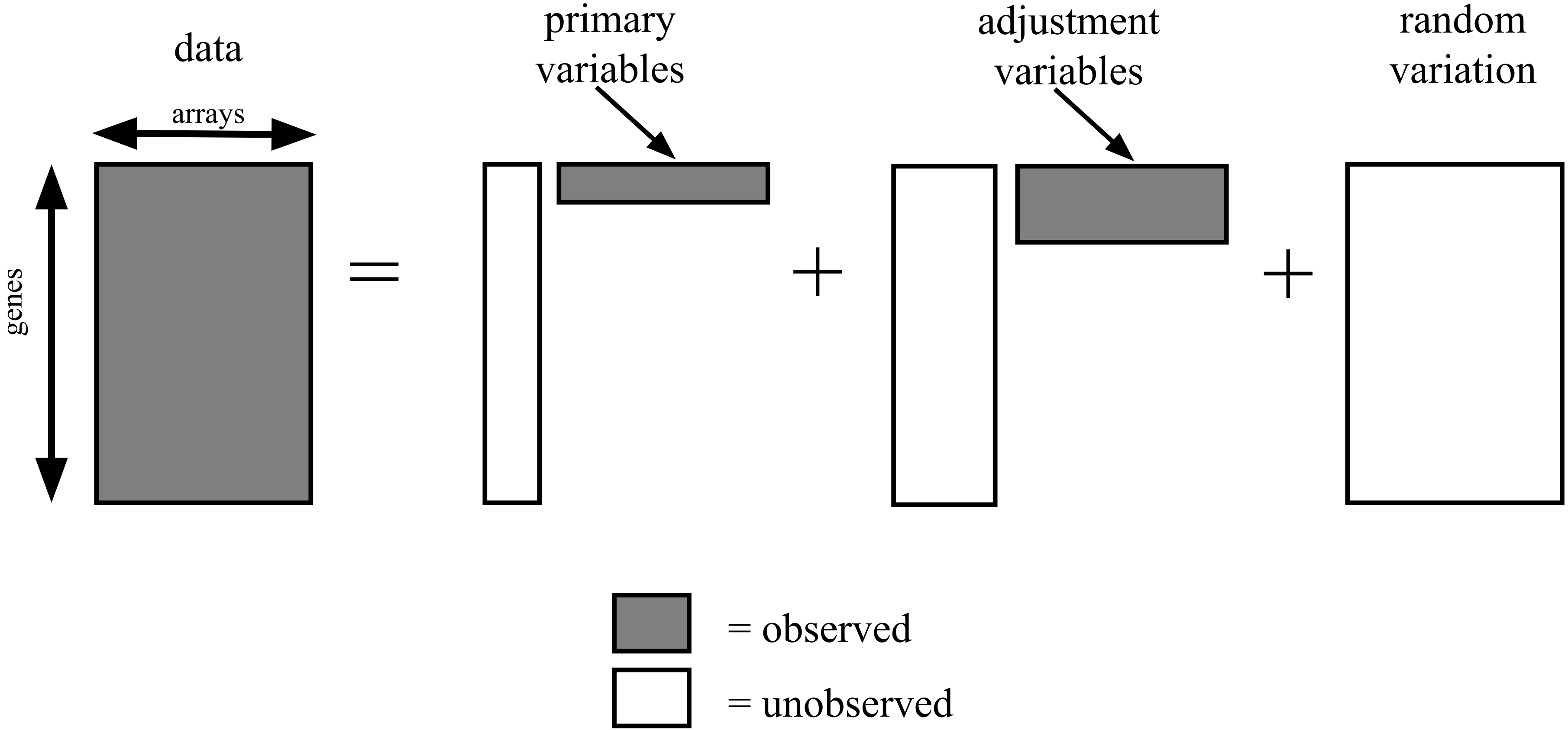
1	**1**	**1**	**1**	**…**	**1**	**1**	**1**
M	F	M	M	…	F	F	M
A	B	A	B	…	B	A	B

Slide adapted from: John Storey

- Obvious:
  - Location (Agadir, Desert, Village)
  - Sex (female, male)
  - Batch (data generated in two batches)
- Subtle:
  - Intensity dependent effects
  - Dye effects
  - Probe composition effect

- Unknowns???

Slide adapted from: John Storey



data

arrays

genes

=

primary variables

+

adjustment variables

+

random variation

= observed

= unobserved

Result

- A set of hundreds, thousands, or millions of model fits
- For each model we have estimates, residuals, fitted values.
- There can now be structure in the estimates, structure in the noise, and all sorts of other issues (we’ll address some shortly)

Notes and further reading

- Linear models is a whole class (no joke): https://www.coursera.org/course/regmods
- Linear models for microarray data: http://www.statsci.org/smyth/pubs/limma-biocbook-reprint.pdf