Acknowledgment

This research was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and the joint National Science Foundation / NIGMS Mathematical Biology Program under award number R01GM109458. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

- DNA encodes all biological information

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

- DNA encodes all biological information
- regions of DNA (genes) encode blueprints for proteins

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

- DNA encodes all biological information
- regions of DNA (genes) encode blueprints for proteins
- messenger RNA (mRNA) conveys information to ribosomes

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

- DNA encodes all biological information
- regions of DNA (genes) encode blueprints for proteins
- messenger RNA (mRNA) conveys information to ribosomes
- ribosomes assemble proteins

Gene expression

(Datta & Nettleton 2014, *Statistical Analysis of Next Generation Sequencing Data*)

- DNA encodes all biological information
- regions of DNA (genes) encode blueprints for proteins
- messenger RNA (mRNA) conveys information to ribosomes
- ribosomes assemble proteins

Gene expression is regulated by mRNA.

Gene expression profiling - RNA-seq

Gene expression profiling - RNA-seq

RNA-seq is:

Gene expression profiling - RNA-seq

RNA-seq is:

- "whole genome shotgun sequencing"

Gene expression profiling - RNA-seq

RNA-seq is:

- "whole genome shotgun sequencing"
- simultaneously measurement of transcript abundance of thousands of genes at once

Gene expression profiling - RNA-seq

RNA-seq is:

- "whole genome shotgun sequencing"
- simultaneously measurement of transcript abundance of thousands of genes at once
- ... at single-base resolution

Steps to producing RNA-seq data:

Gene expression profiling - RNA-seq

RNA-seq is:

- "whole genome shotgun sequencing"
- simultaneously measurement of transcript abundance of thousands of genes at once
- ... at single-base resolution

Steps to producing RNA-seq data:

- isolate mRNA and fragment it

Gene expression profiling - RNA-seq

RNA-seq is:

- ”whole genome shotgun sequencing”
- simultaneously measurement of transcript abundance of thousands of genes at once
- ... at single-base resolution

Steps to producing RNA-seq data:

- isolate mRNA and fragment it
- match fragments back to genes/features (counts)

Motivating example

Heterosis

- “Hybrid vigor”

Motivating example

Heterosis

- “Hybrid vigor”
- Larger offspring than parents

Motivating example

Heterosis

- “Hybrid vigor”
- Larger offspring than parents
- Complementation

Motivating example

Heterosis

- "Hybrid vigor"
- Larger offspring than parents
- Complementation
- High parent heterosis (HPH) $LP < HP < H$

Motivating example

Heterosis

- "Hybrid vigor"
- Larger offspring than parents
- Complementation
- High parent heterosis (HPH) $LP < HP < H$
- Low parent heterosis (LPH) $H < LP < HP$

RNA-seq data

	Population 1			Population 2		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Gene 1	7	2	0	14	18	41
Gene 2	55	42	40	32	22	37
Gene 3	41	40	32	61	61	60
Gene 4	40	43	35	15	24	39
raw library size	11569434	10079799	9028465	10028258	9010306	10283594

Data from Paschold et al. (2012)

Research goal: To use gene expression data to identify genes responsible for hybrid vigor (heterosis).

Data from Paschold et al. (2012)

Research goal: To use gene expression data to identify genes responsible for hybrid vigor (heterosis).

- 2 recombinant inbred lines (homozygous): B73, Mo17
- 2 reciprocal hybrid crosses: B73×Mo17, Mo17×B73
- 4 replicates of each variety
- sequencing done on 2 flow cells, replicates balanced across flow cells

Normalization

Definitions

r_{gn} = count for gene g, sample n; $\quad g = 1, \dots, G, \quad n = 1, \dots, N$

R_n = library size for sample n

$$\log \tilde{R} = \frac{1}{N} \sum_{n=1}^{N} \log(R_n)$$

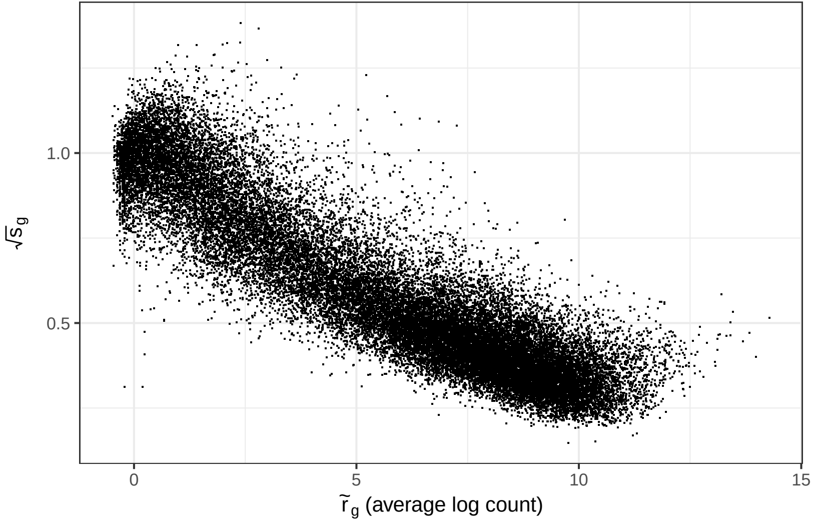
$$y_{gn} = \log_2 \left(\frac{r_{gn} + 0.5}{R_n + 1} \times 10^6 \right)$$

$$\tilde{r}_g = \frac{1}{N} \sum_{n=1}^{N} y_{gn} + \log_2(\tilde{R}) - \log_2(10^6)$$

X = model matrix

$s_g = \sqrt{MSE}$ from fitting a linear model, $y_g \sim N(X\beta_g, \sigma_g^2 I_n)$

Mean-variance in RNA-seq


$\sqrt{s_g}$
$\tilde{r}_g$ (average log count)