Some challenges related to next-generation sequencing data

Rafael A. Irizarry
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

http://rafalab.org
@rafalab
http://genomics.jhu.edu

Topics

- Base calling
- Biological variation and systematic biases
- Batch effects (Time permitting)
- Bump hunting (offline)

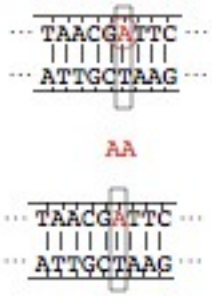
Base calling most slides courtesy of

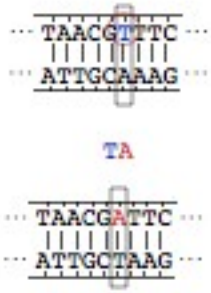
Héctor Corrada Bravo and Joyce Hsiao
Center for Bioinformatics and Computational Biology
Dept. of Computer Science
University of Maryland-College Park

http://cbcb.umd.edu/~hcorrada

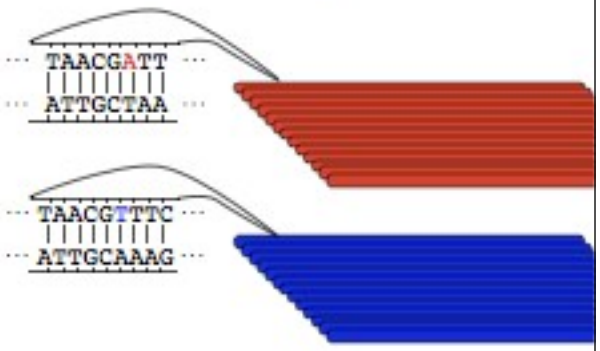
What is the basis of phenotypic variation?

Genotyping

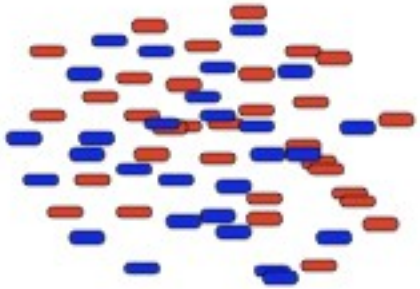




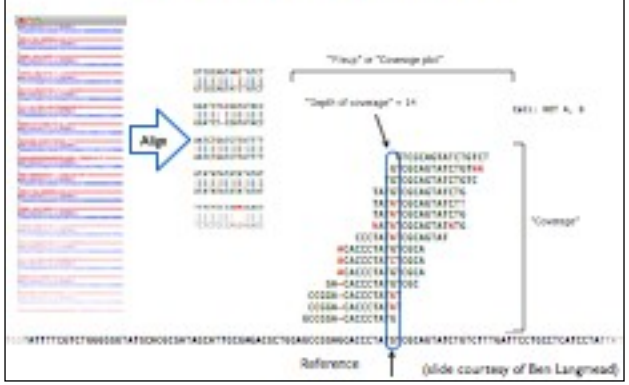
Sec-gen Sequencing for SNPs



Sec-gen Sequencing for SNPs



Sec-gen Sequencing for SNPs

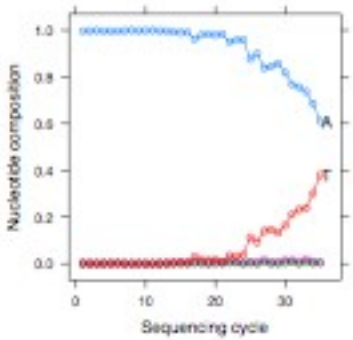


(slide courtesy of Ben Langmead)

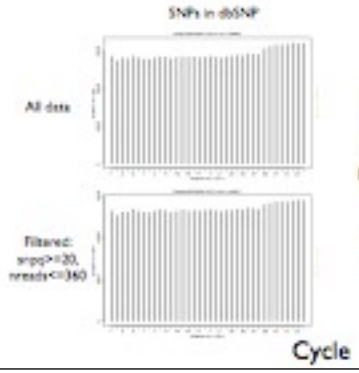
SNPs

SNPs

All Reads

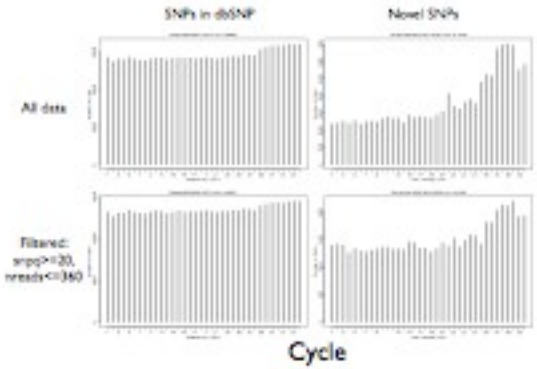


1000 Genomes Data

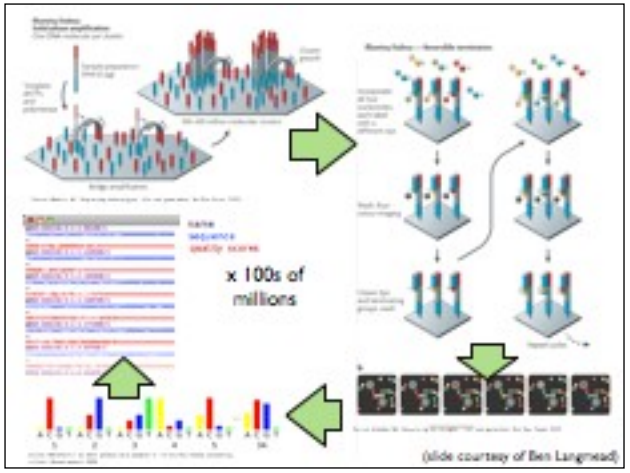


Here we aggregate reads and record cycle at which variant appears

1000 Genomes Data



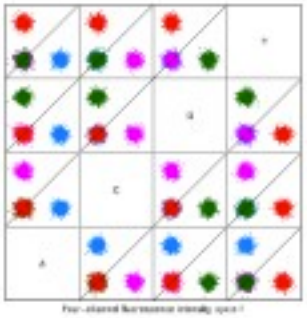
What is causing this?



Before Reads There were Intensities

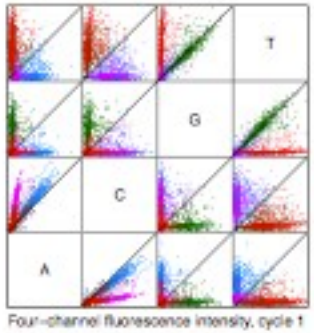
```
> head(I[1:10,1:4])
     A.1    C.1    G.1    T.1
1  154.8  122.1  119.3 13881.9
2 1893.5 6186.6 -798.4   288.3
3  892.3 4859.2 -367.9  -403.0
4  398.5 2587.9  -85.6   188.7
5  979.6 6411.8  943.5  -454.9
6  945.5 6963.1   18.7 -1178.8
7  233.8  213.3   15.5  4358.8
8 1885.2 5834.5 -384.7   -94.1
9  267.6  348.3 6866.2  5788.6
10 1162.6 8424.4 -497.6  -149.2
```

We Want to See This



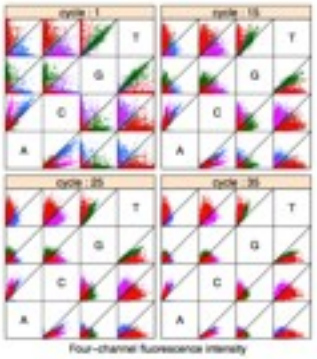
Color coded by call made: A, C, G, T

But See This

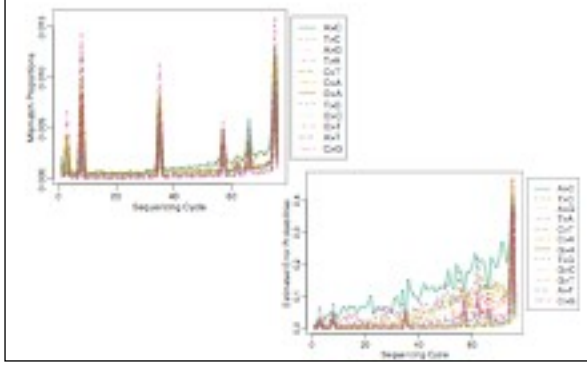


Color coded by call made: A, C, G, T

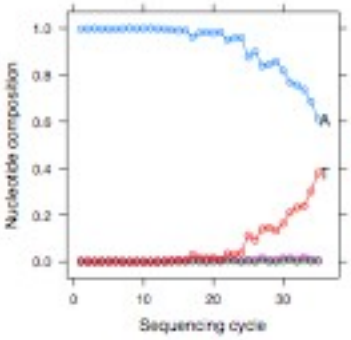
Gets Worse for higher cycles



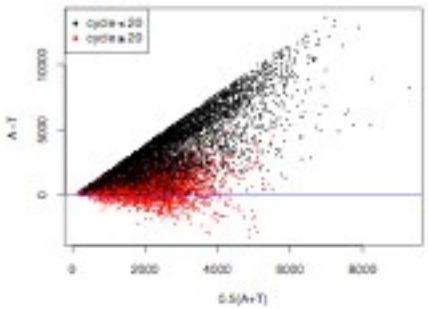
Error Rate and Reported Quality



Remember This?



Bias Explained



Base Calling

1) Rougemont et al. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics (2008)

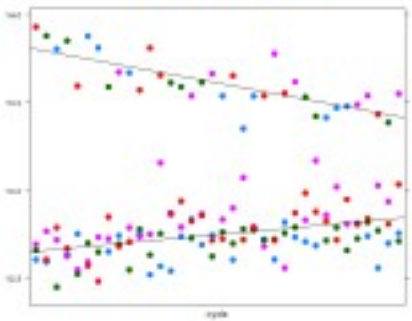
2) Erlich et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods (2008)

3) Kao et al. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res (2009)

4) Cornuéjols Bravo and Irizarry. Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data. Biometrics (2009)

5) Cokus et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature (2009)

Intensity Model



Intensity Model

log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \underline{\Delta_{ijc}}(\mu_{cj\alpha} + x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}) + \underline{(1 - \Delta_{ijc})}(\mu_{cj\beta} + x_j^T \beta_i + \epsilon_{ijc}^{\beta})$$

indicators of nucleotide identity, read *i*, pos. *j*

$$\Delta_{ijc} = \begin{cases} 1 & \text{if } c \text{ is the nucleotide in read } i \text{ position } j \\ 0 & \text{otherwise} \end{cases}$$

Intensity Model

log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + \underline{x_j^T \alpha_i} + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(\mu_{cj\beta} + \underline{x_j^T \beta_i} + \epsilon_{ijc}^{\beta})$$

read-specific linear models

Intensity Model

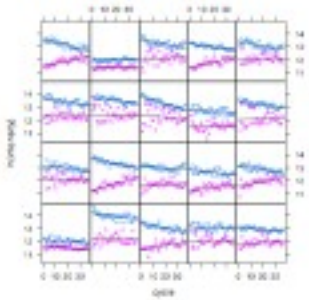
log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + x_j^T \alpha_i + \underline{\epsilon_{ijc}^{\alpha}}) + (1 - \Delta_{ijc})(\mu_{cj\beta} + x_j^T \beta_i + \underline{\epsilon_{ijc}^{\beta}})$$

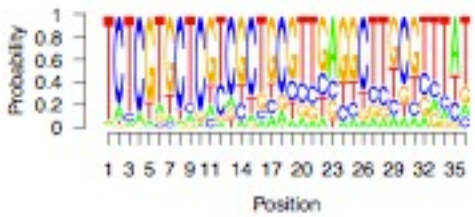
measurement error

$$\epsilon_{ijc}^{\alpha} \sim N(0, \sigma_{\alpha i}^2) \qquad \epsilon_{ijc}^{\beta} \sim N(0, \sigma_{\beta i}^2)$$

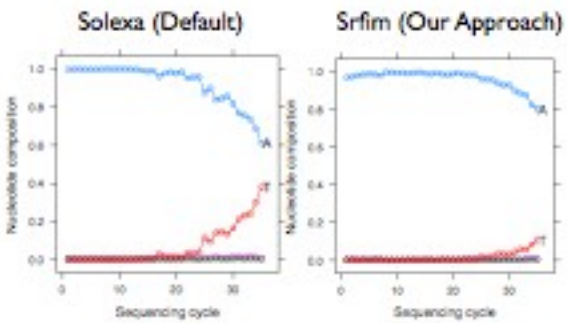
Read & Cycle Effects



Base Identity Probability Profiles



Before And After



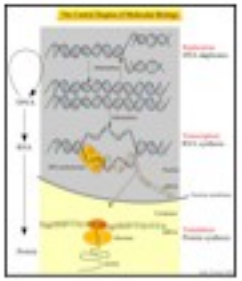
Biological variation and systematic biases

and its importance in gene expression

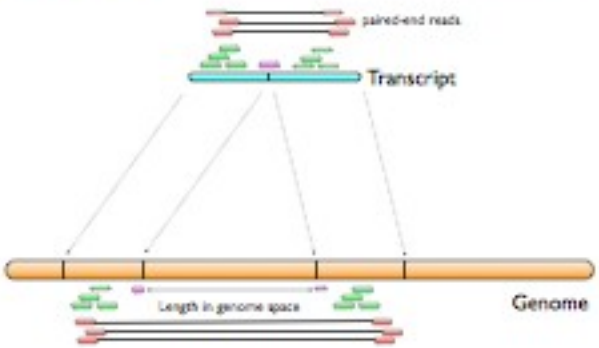
Examples from RNAseq data analysis

What is the basis of phenotypic variation?

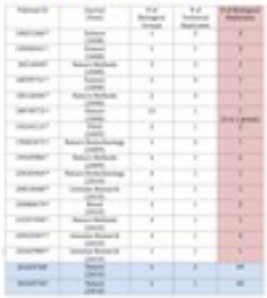
Central Dogma of Biology



Mapping transcripts

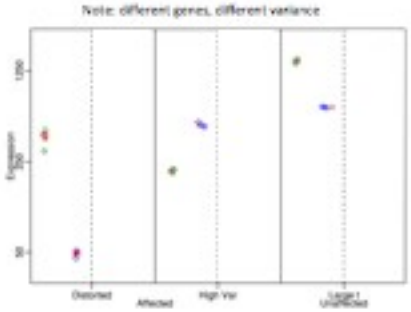


Most RNAseq studies do not include biological replicates



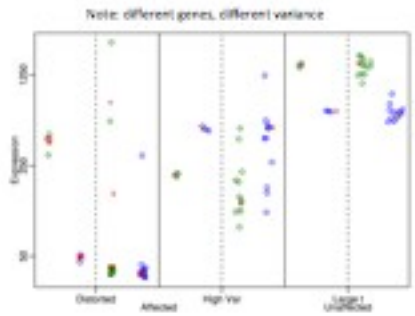
Hansen et al. (2011) *Nature Biotechnology*

Biological versus technical replicates



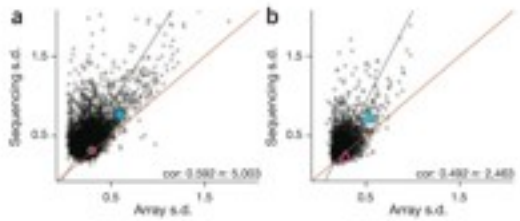
Kendziorski et al. (2005) *PNAS*

Biological versus technical replicates

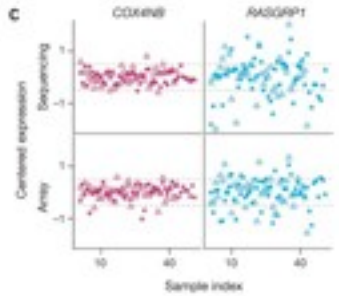


Kendziorski et al. (2005) *PNAS*

SD NGS versus microarray



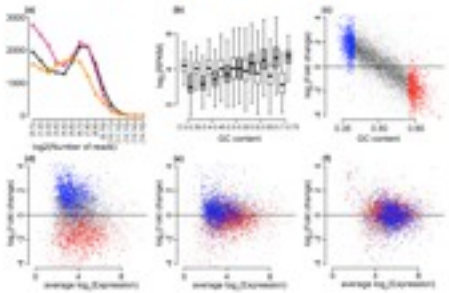
Highlighted genes by sample



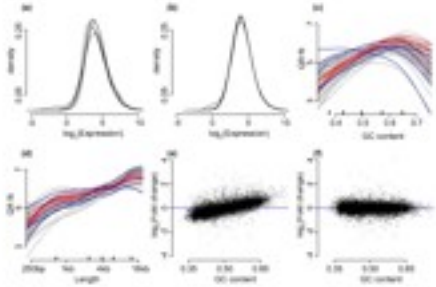
Systematic errors

Plots courtesy of Kasper Hansen and Zhijin Wu

Different distributions and GC content effects

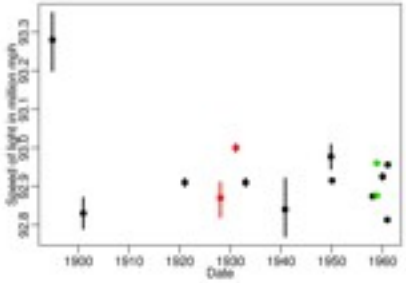


Estimated distributions and GC effects



Batch Effects

Speed of Light Estimates with "Confidence Intervals" (1900-1960)



Youden, WJ (1972), "Enduring Values," Technometrics , 14, 1-11.

Batch Effect in Genomics

Many slides provided by Jeff Leek

Gene Expression

LETTER

Natural variation in human gene expression assessed in lymphoblastoid cells

Gene Expression

LETTER

Natural variation in human gene expression assessed in lymphoblastoid cells

Correspondence

Nature Genetics **39**, 807 - 808 (2007)
doi:10.1038/ng0707-807

On the design and analysis of gene expression studies in human populations

Joshua M Akey[1], Shameek Biswas[1], Jeffrey T Leek[1] & John D Storey[1,2]

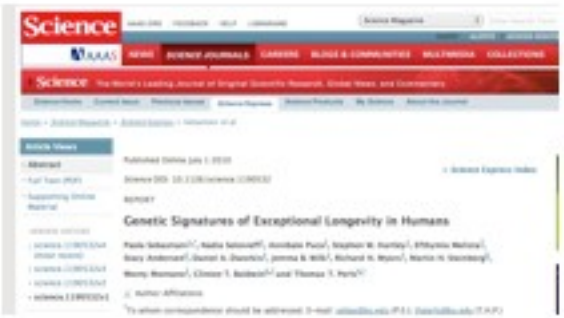
Proteomics

THE LANCET

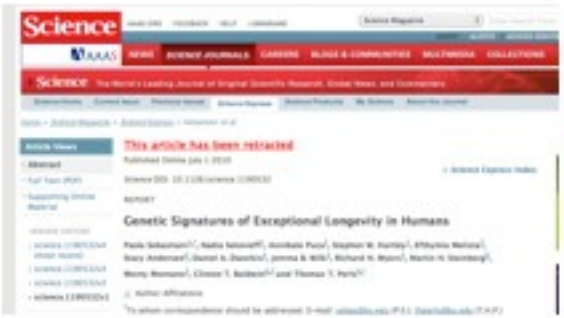
Use of proteomic patterns in serum to identify ovarian cancer

Proteomics

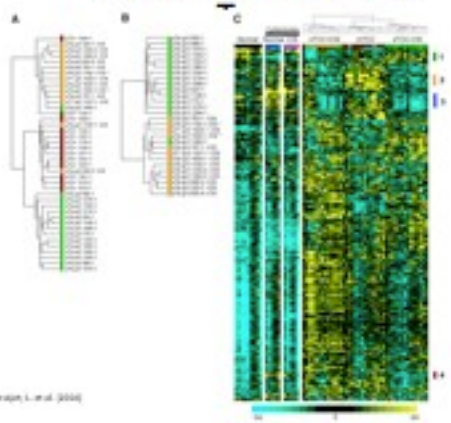
GWAS



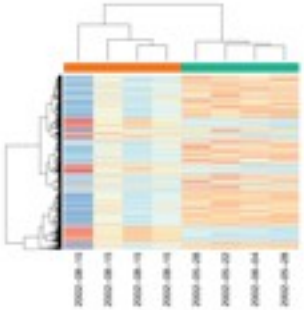
GWAS



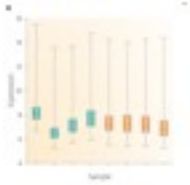
Where does it come from?



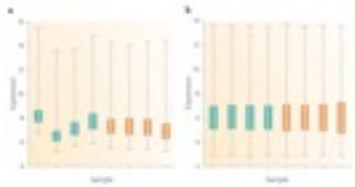
Clustering of normals



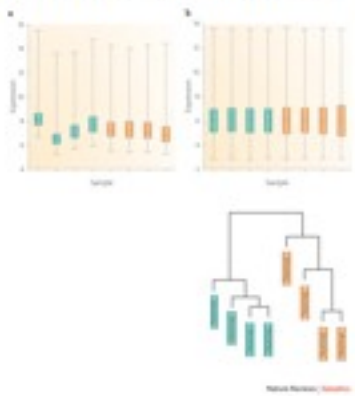
What are Batch Effects



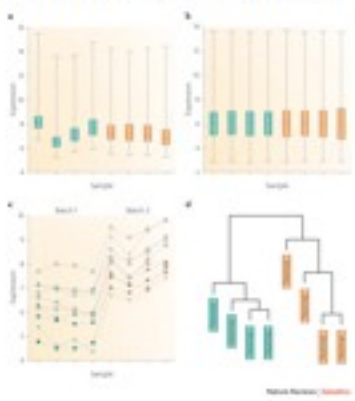
What are Batch Effects



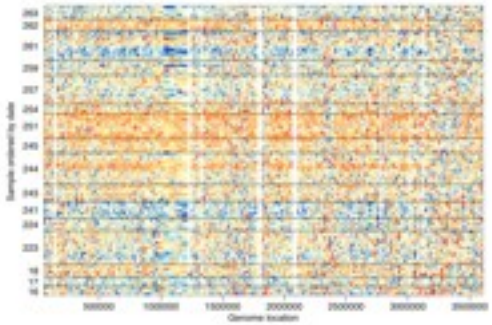
What are Batch Effects



What are Batch Effects



Batch Effects in Sequencing



Math gives us hope

NATURE REVIEWS | GENETICS

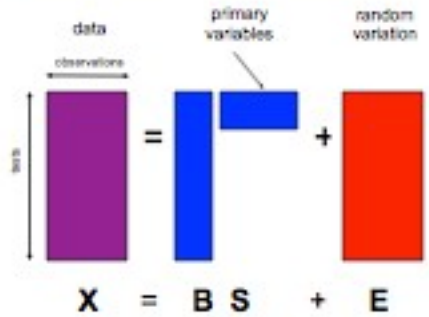
OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

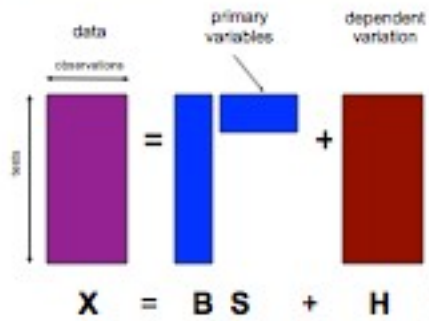
Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Thanks

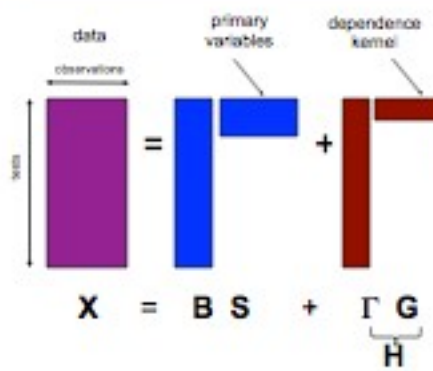
Decomposing Heterogeneity

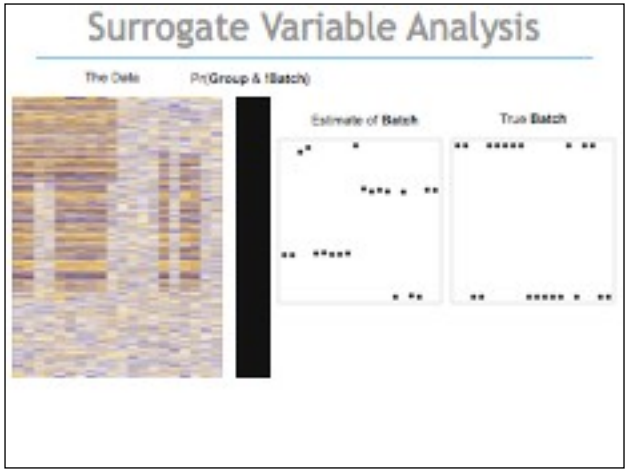


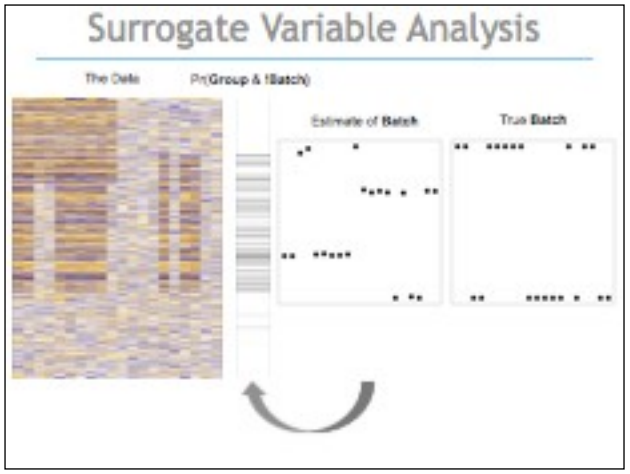
Decomposing Heterogeneity

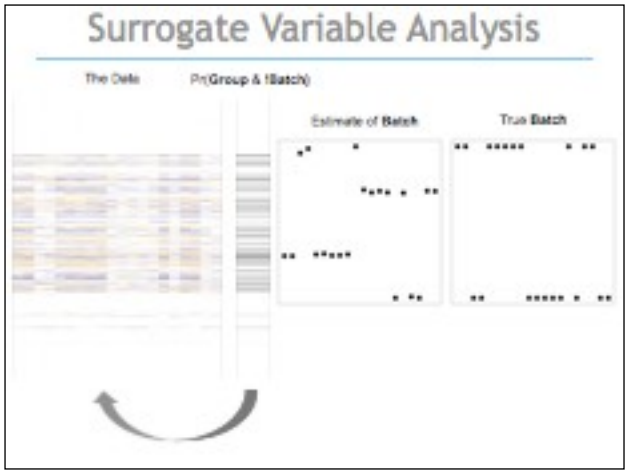


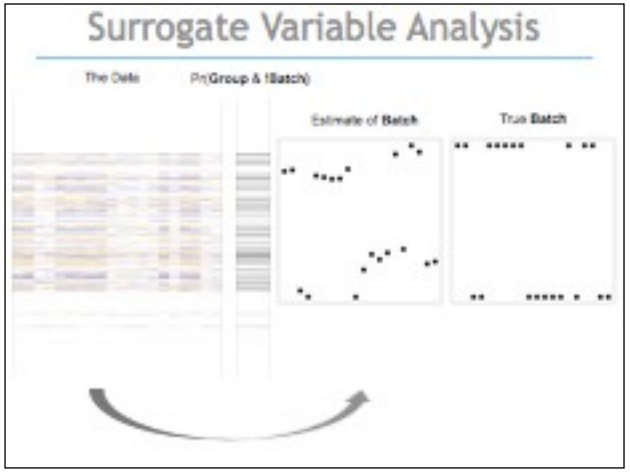
Decomposing Heterogeneity

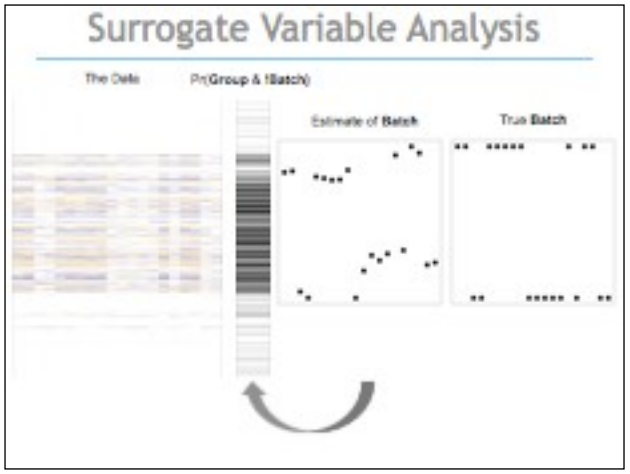


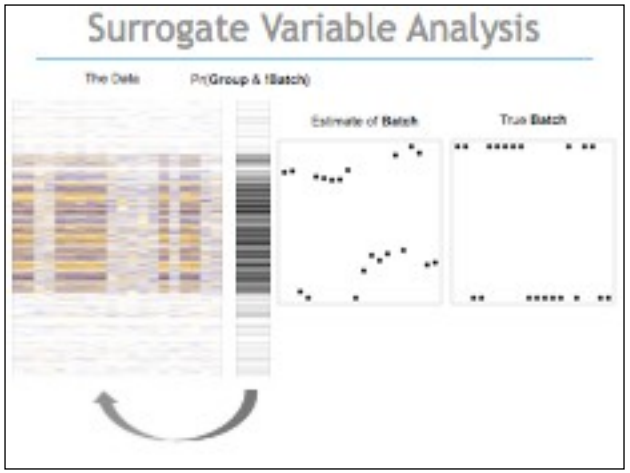

Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch


Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch

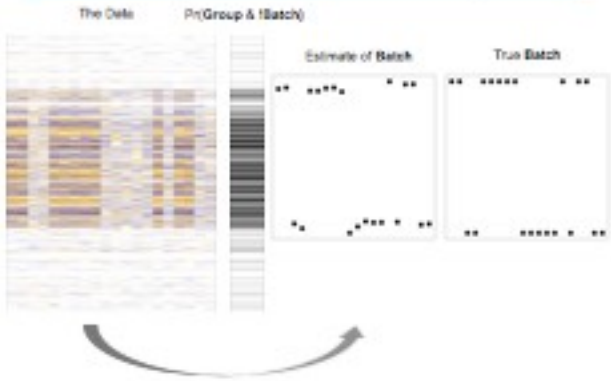

Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch


Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch

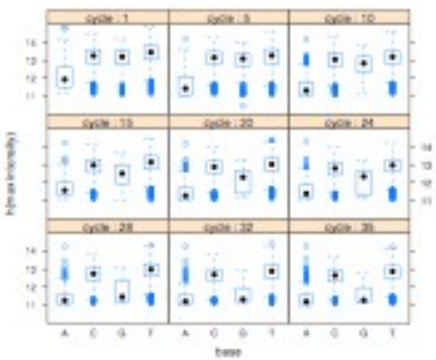

Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch


Surrogate Variable Analysis
The Data
Pr(Group & !Batch)
Estimate of Batch
True Batch

Surrogate Variable Analysis



Base-Cycle Effect



Intensity Model

log intensity read i, cycle j, channel c

$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(\mu_{cj\beta} + x_j^T \beta_i + \epsilon_{ijc}^{\beta})$$

estimate parameters with EM algorithm, which also estimates

$$z_{ijc} := \mathrm{E}\{\Delta_{ijc} = 1 | u_{ij}\} = P(\Delta_{ijc} = 1 | u_{ij})$$

Resources

Papers:

- Corrada Bravo & Irizarry, Biometrics, November 2010
- Wu, Irizarry & Corrada Bravo, Nat. Methods, May 2010
- Niranjan, et al., Genome Biology, Sept. 2011.

Software: http://cbcb.umd.edu/~hcorrada/seqgen

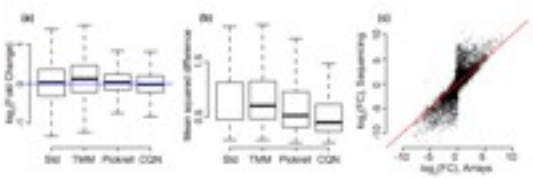
- Srfim: Illumina basecalling
- Rsolid: Preprocessing for SOLiD
- Service: Variant calling in targeted pooled samples

Model

$$Y_{g,i} \mid \mu_{g,i} \sim \text{Poisson}(\mu_{g,i})$$

$$\mu_{g,i} = \exp\left\{ h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) \right\}$$

Results

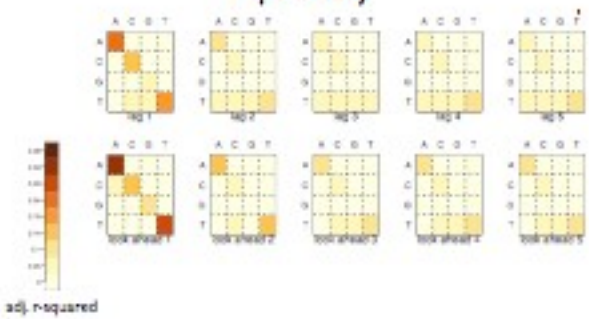


The trick

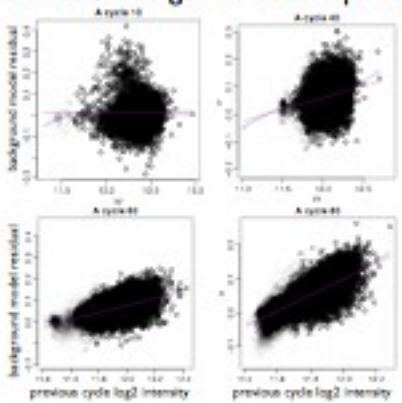
$$\log(Y_{g,i}) \mid \mu_{g,i} \approx \log(\mu_{g,i}) = h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}),$$

$$\mathrm{E}\left\{\log(Y_{g,i}) - \sum_{j=1}^{p} f_{i,j}(X_{g,j})\right\} = h_i(\theta_{g,i})$$

Cross-cycle and cross-channel dependency



The next challenge:Residual Dependency



Residual Dependency