Understanding and Validating Experimental Expectations

Festival of Genomics 2017

Simon Andrews
simon.andrews@babraham.ac.uk

Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Types of Expectation

- Nature of samples — Human Male Liver
- Nature of data — RNA-Seq/Genomic
- Efficacy of processing — Equal losses

- Effect of interventions — Did they work
- Nature of effects — Global/Local
- Sources of variation — Any unexpected

Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Raw Data Expectation

ENVIRONMENTAL Science & Technology

ACS AuthorChoice

Article

pubs.acs.org/est

Tris(1,3-dichloro-2-propyl)phosphate Induces Genome-Wide Hypomethylation within Early Zebrafish Embryos

TDCIPP exposure predominantly resulted in hypomethylation of positions outside of CpG islands and within intragenic (exon) regions of the zebrafish genome.

[†]Department of Environmental Sciences, University of California, Riverside, Riverside, California 92521, United States

[‡]Department of Environmental Health Sciences, University of South Carolina, Columbia, South Carolina 29208, United States

[§]Division of Environmental Sciences and Policy, Duke University, Durham, North Carolina 27708, United States

[∥]Bioinformatics Facility, Institute for Integrative Genome Biology, University of California, Riverside, Riverside, California 92521, United States

[⊥]Department of Chemistry, University of California, Riverside, Riverside, California 92521, United States

Supporting Information

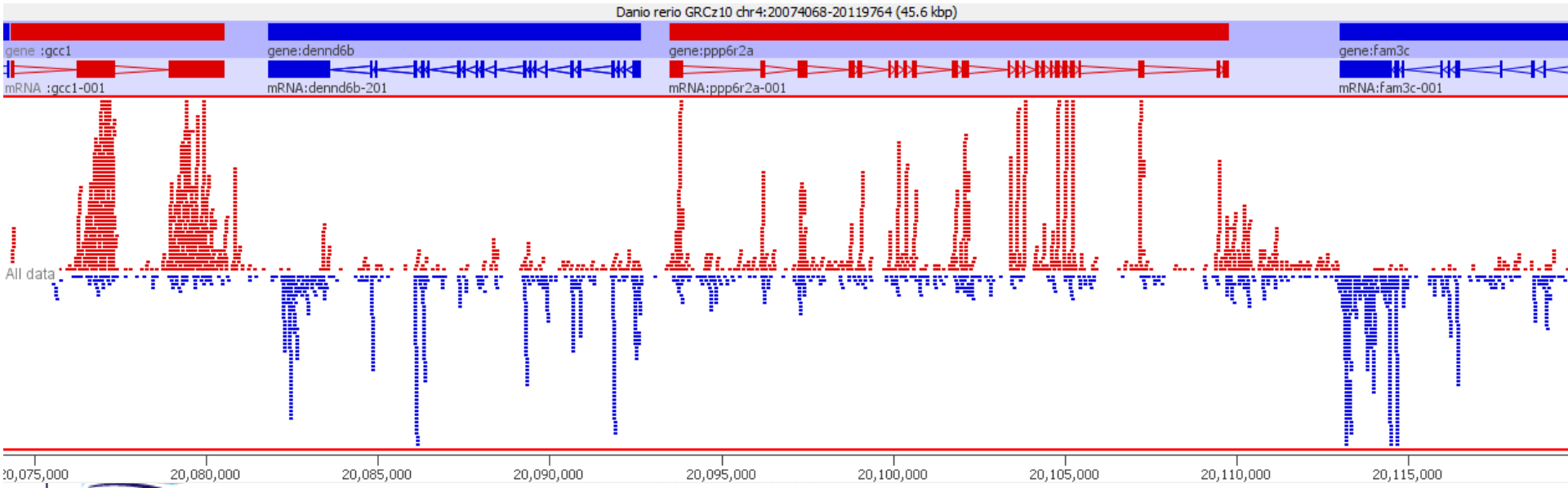
Babraham Institute

UNIVERSITY OF CAMBRIDGE

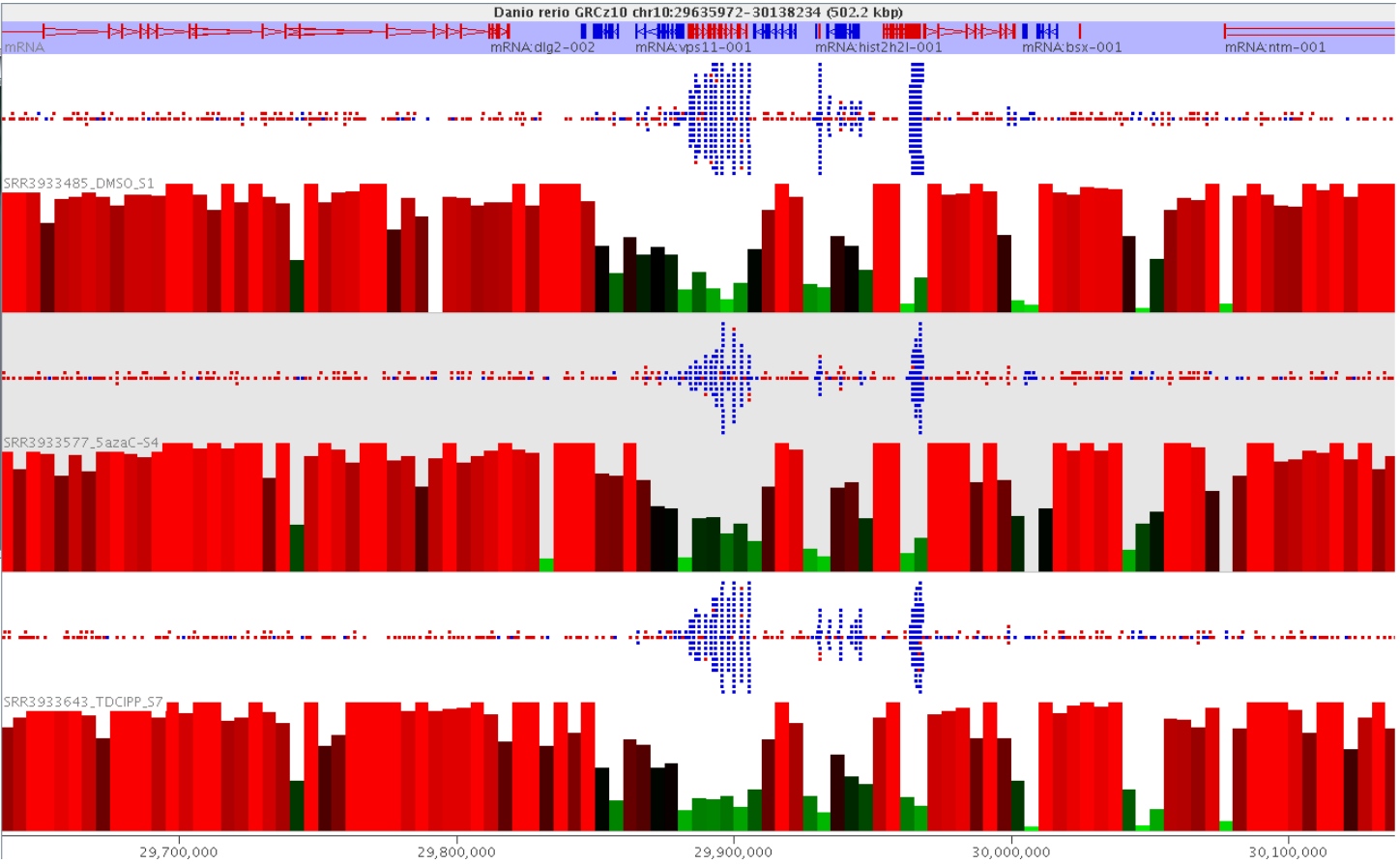
Babraham Bioinformatics

Raw Data Expectations

- Bisulphite Sequencing
  - Whole genome – all regions equally sampled
  - Both strands – no read level strand bias



RNA-Contamination



Calls
Red = meth
Blue = unmeth

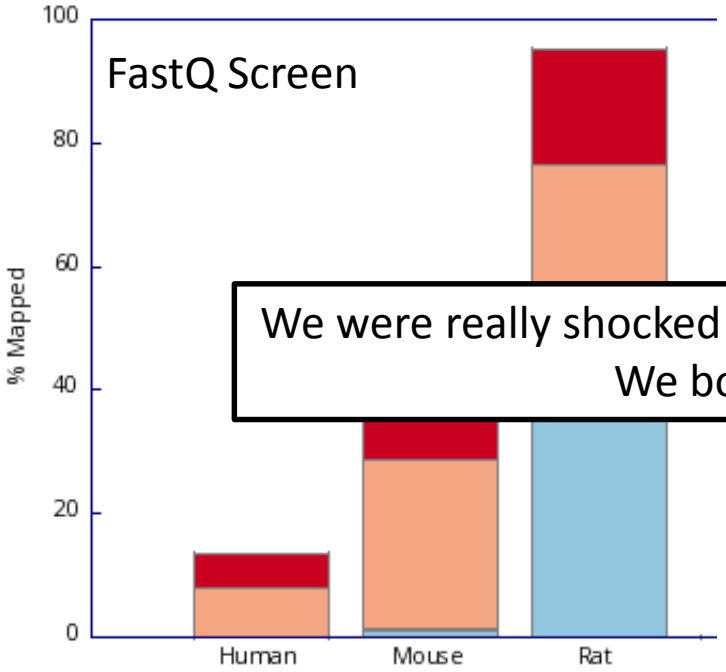
Methylation level

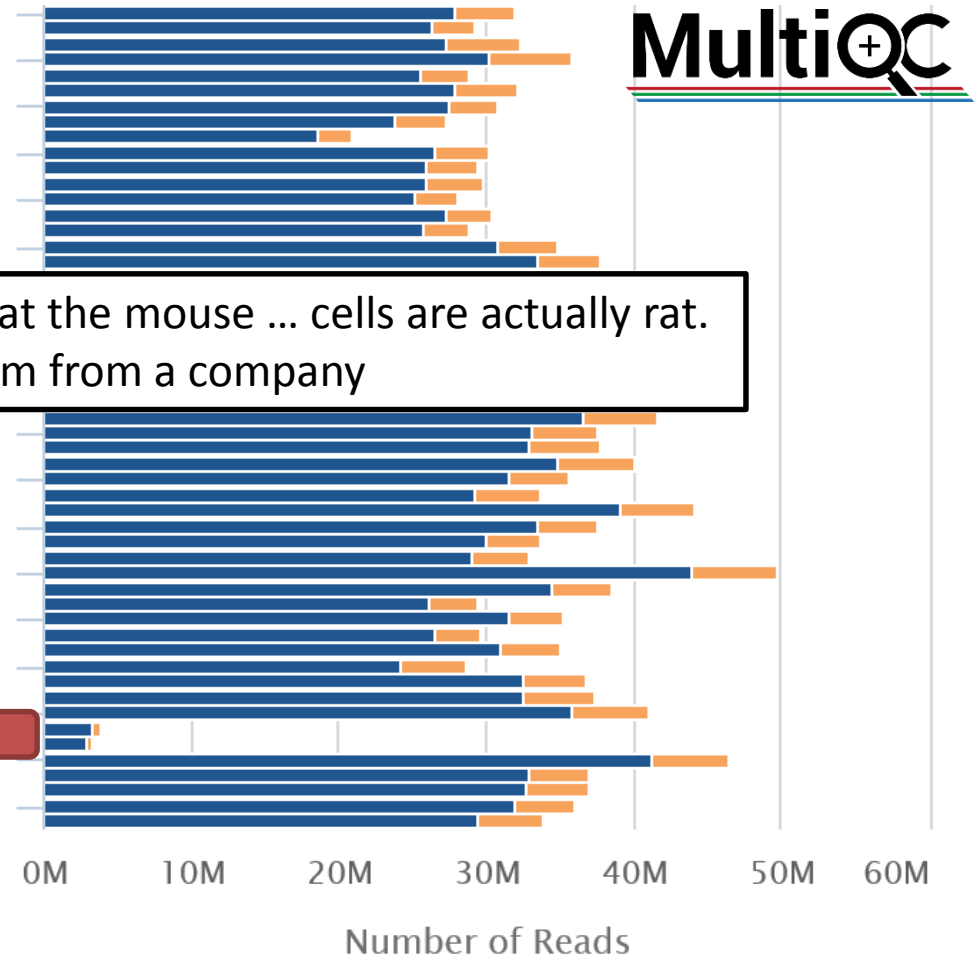
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Processing Expectations (Mouse RNA-Seq)





We were really shocked to see that the mouse … cells are actually rat. We bought them from a company

Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Expectations

Your analysis plan is intrinsically linked to your expectations

analysis *data*
"No ~~battle~~ plan survives contact with the ~~enemy.~~"

Helmuth von Moltke

Babraham Institute

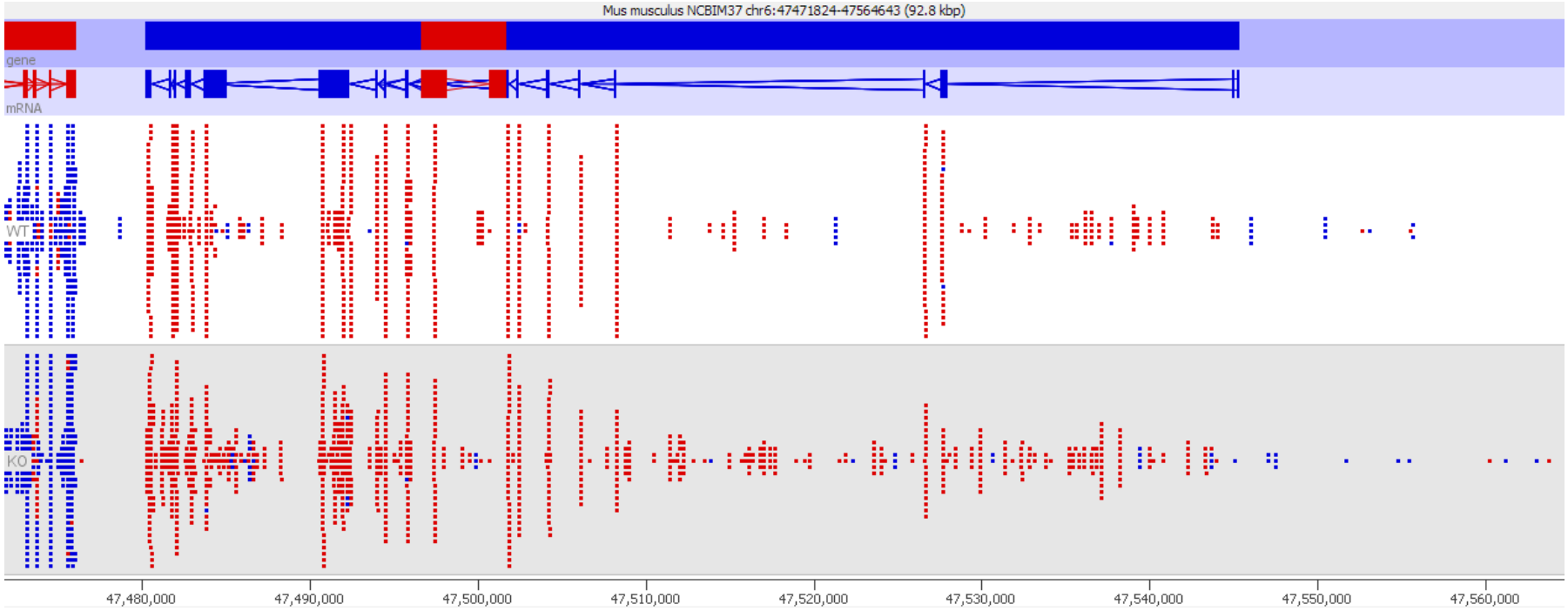
UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Gene KO Biological Assumptions

- The knockout experimental strategy worked as expected
- The reduction in transcript is large enough to achieve a biological effect
- The system didn't find a simple way to compensate

Expected Effects



Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Compensation

cells

MDPI

Article

Random Splicing of Several Exons Caused by a Single Base Change in the Target Exon of CRISPR/Cas9 Mediated Gene Knockout

Marcel Kapahnke, Antje Banning and Ritva Tikkanen *

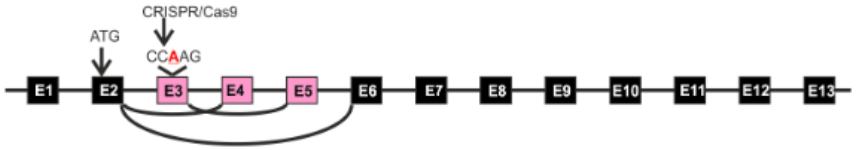
Institute of Biochemistry, Medical Faculty, University of Giessen, Friedrichstrasse 24, D-35392 Giessen, Germany; Marcel.Kapahnke@med.uni-giessen.de (M.K.); Antje.Banning@biochemie.med.uni-giessen.de (A.B.)
* Correspondence: Ritva.Tikkanen@biochemie.med.uni-giessen.de; Tel.: +49-641-9947-420

Academic Editor: Alexander E. Kalyuzhny
Received: 20 October 2016; Accepted: 9 December 2016; Published: 14 December 2016

Abstract: The clustered regularly interspaced short palindromic repeats (CRISPR)-associated sequence 9 (CRISPR/Cas9) system is widely used for genome editing purposes as it facilitates an efficient knockout of a specific gene in, e.g. cultured cells. Targeted double-strand breaks are introduced to the target sequence of the guide RNAs, which activates the cellular DNA repair mechanism for non-homologous-end-joining, resulting in unprecise repair and introduction of small deletions or insertions. Due to this, sequence alterations in the coding region of the target gene frequently cause frame-shift mutations, facilitating degradation of the mRNA. We here show that such CRISPR/Cas9-mediated alterations in the target exon may also result in altered splicing of the respective pre-mRNA, most likely due to mutations of splice-regulatory sequences. Using the human *FLOT-1* gene as an example, we demonstrate that such altered splicing products also give rise to aberrant protein products. These may potentially function as dominant-negative proteins and thus interfere with the interpretation of the data generated with these cell lines. Since most researchers only control the consequences of CRISPR knockout at genomic and protein level, our data should encourage to also check the alterations at the mRNA level.

Keywords: RNA splicing; CRISPR; Cas9; genome editing; flotillin

(a) Clone 2



(b) Clone 7

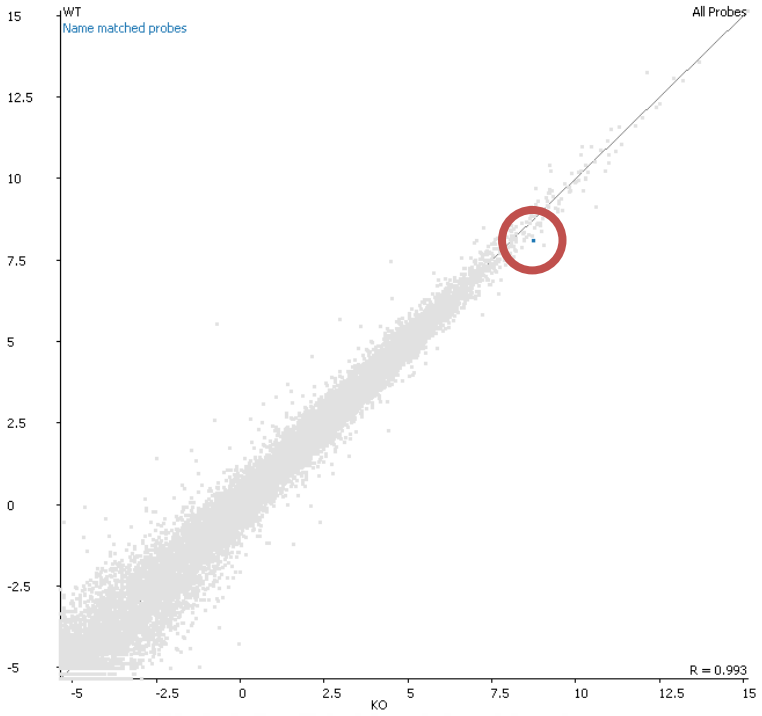
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Biological Relevance

- Heterozygous gene knockout
- Giving very few hits through a standard pipeline



Babraham Institute

UNIVERSITY OF CAMBRIDGE

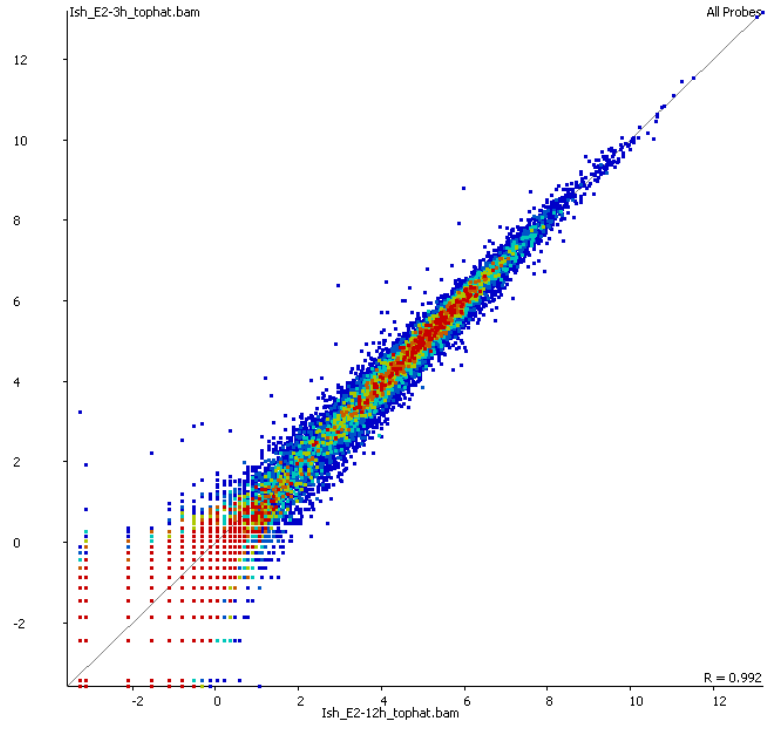
Babraham Bioinformatics

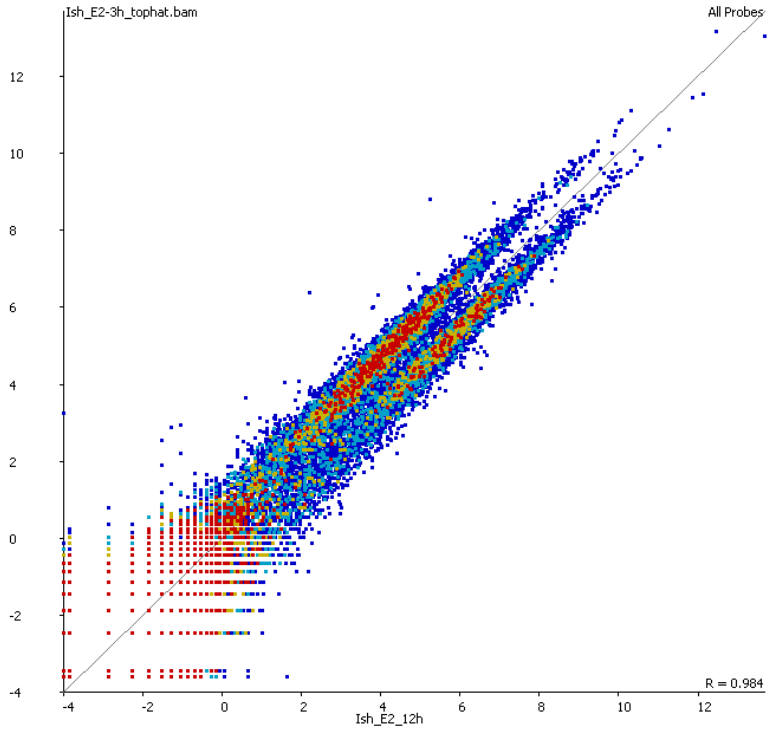
Expected Changes Assumptions

- The change will only directly affect a limited subset of genes
- Genes which are highly affected by the change will be split between being downregulated and upregulated
- The general patterning of transcript expression will not change
- The change will be similar in all biological replicates

Quantitations come with Assumptions

Standard Log2 Reads per Million Reads of Library Quantitation





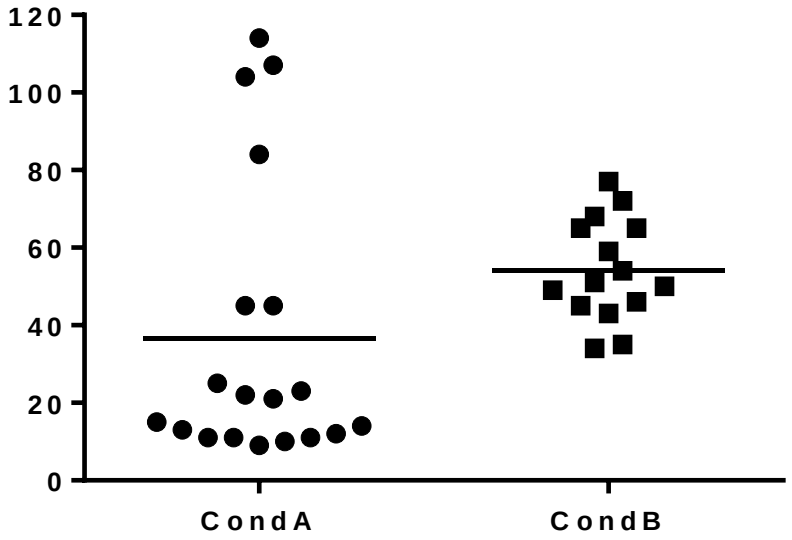
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Statistics come with Assumptions

- T-test
  - Data is normally distributed
  - Variances are equal
  - Replicates are consistent



Babraham Institute

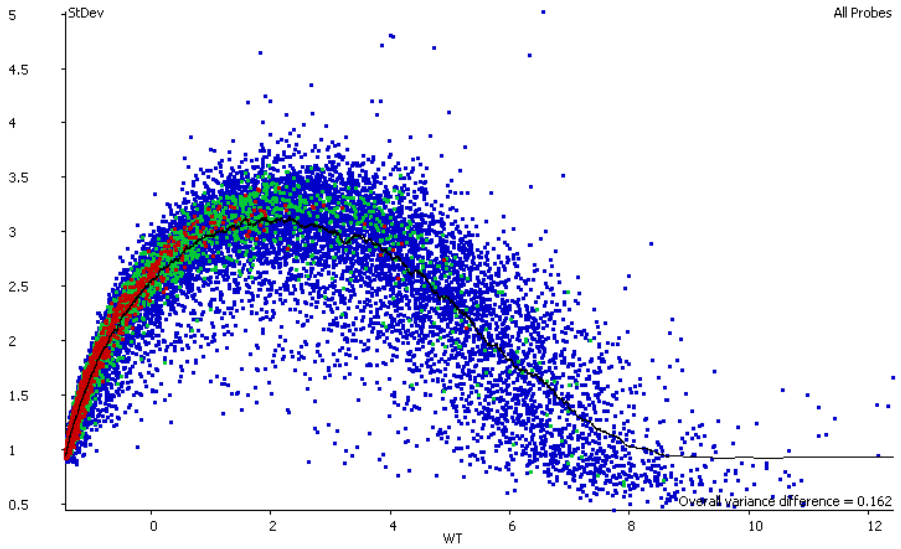
UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Statistics come with Assumptions

DESeq / EdgeR / BaySeq etc

Use variance information sharing between genes with similar expression levels on the assumption that they will exhibit similar variance

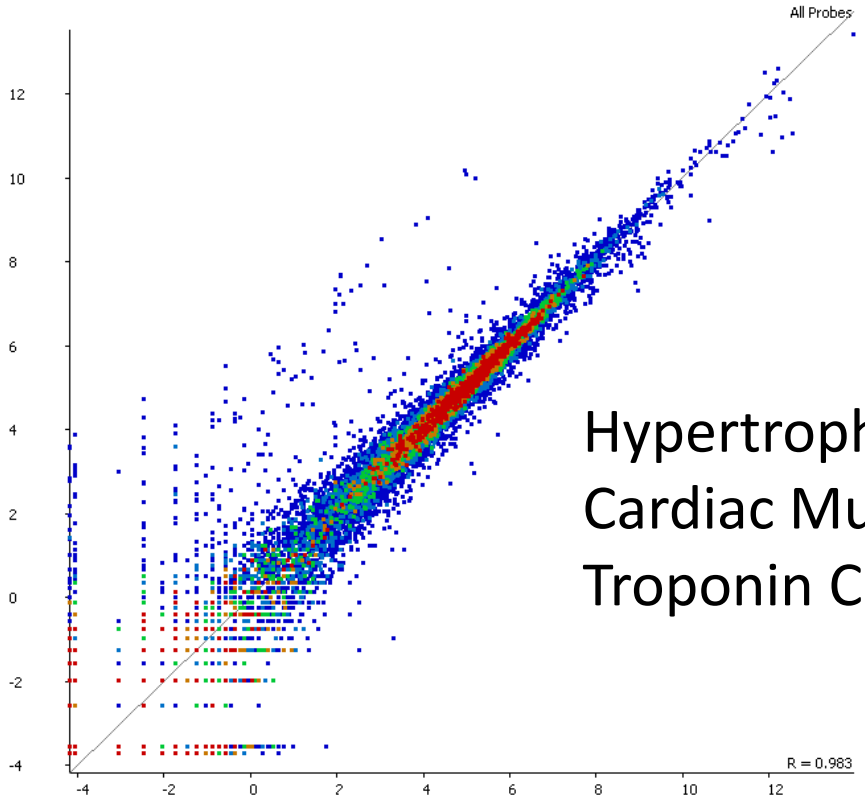


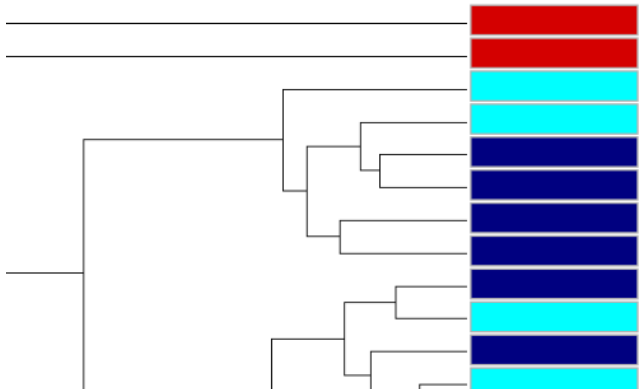
Babraham Institute

UNIVERSITY OF CAMBRIDGE

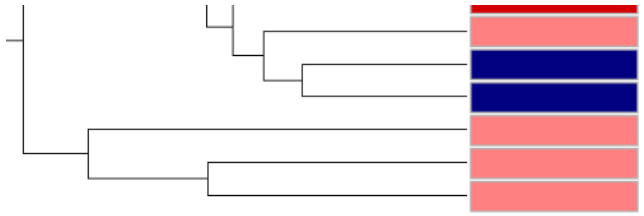
Babraham Bioinformatics

Secondary Signals





Hypertrophic cardiomyopathy (p2e-14)
Cardiac Muscle Contraction (p2e-13)
Troponin Complex (p4e-6)



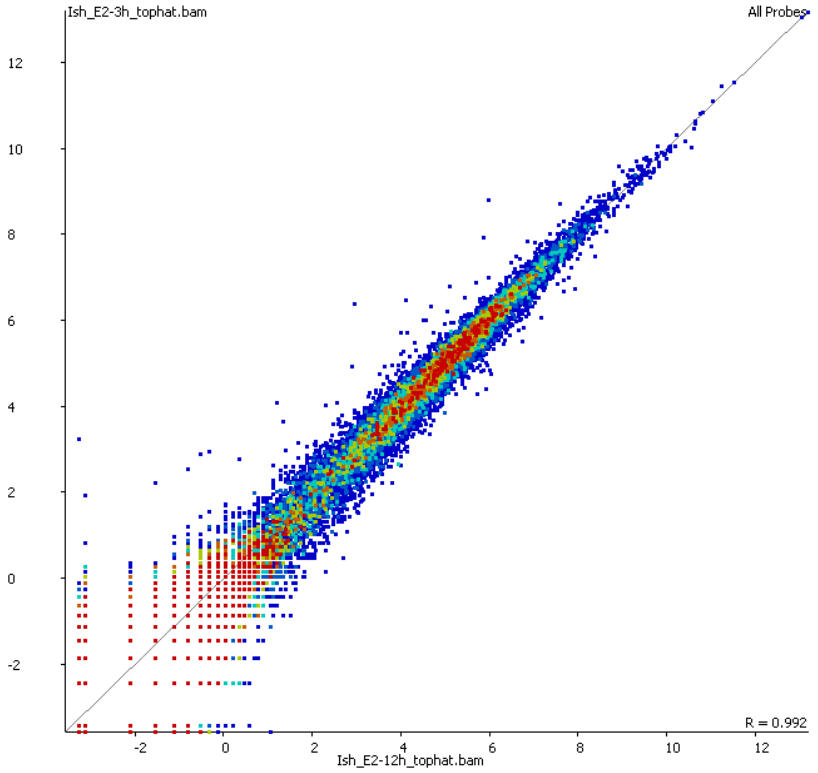
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Make sure you're asking the right question

- Which points change between two conditions?



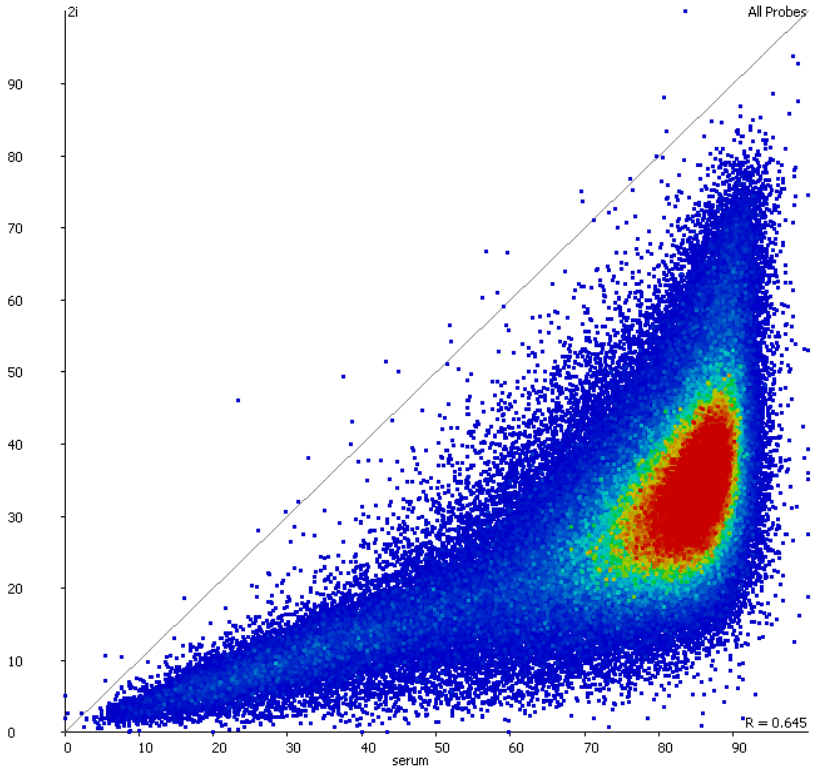
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Make sure you're asking the right question

- Which points change between conditions?
- Which points change between conditions more or less than you'd expect?



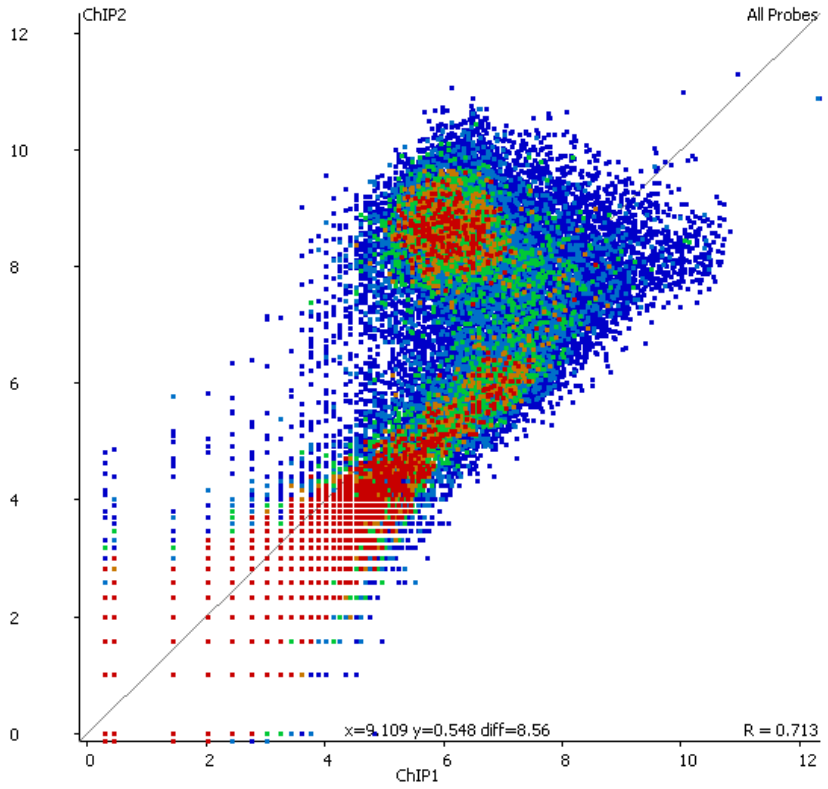
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Make sure you're asking the right question

- Which points change between the two conditions?



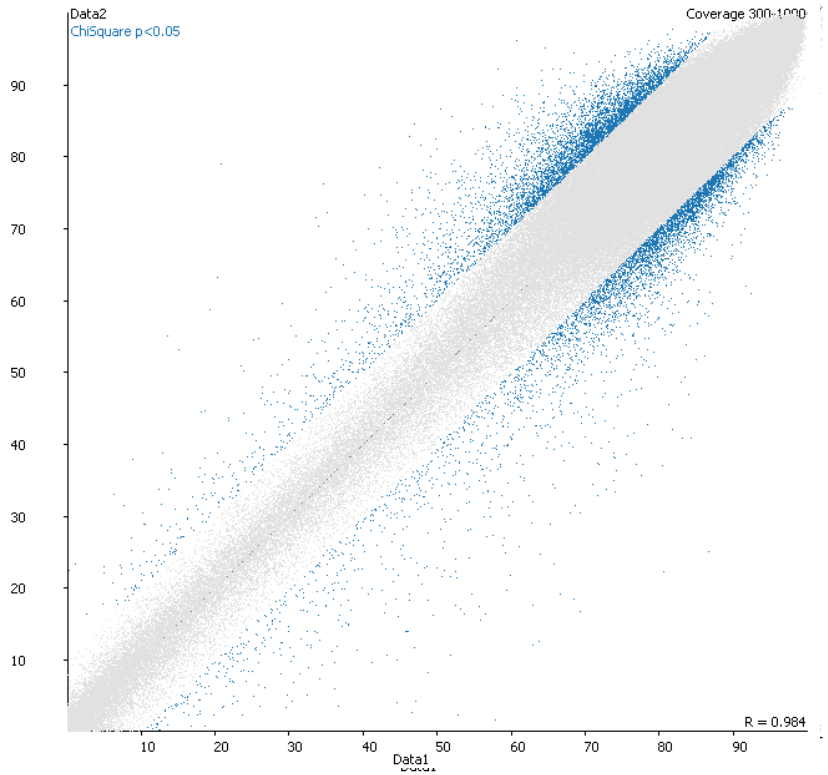
Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Make sure you're asking the right question

- Which points change between two conditions?



Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

What Should We Validate?

- Biological
  - Species
  - Sex
  - Genotype
- Processing
  - Efficiency
  - Types of drop out
  - Categorised results
- Data
  - Genomic distribution
  - Expected effects
  - Sample clustering
  - Overall differences
  - Quantitation
  - Statistical assumptions