Biases and their Effect on Biological Interpretation

Festival of Genomics 2017

Simon Andrews

simon.andrews@babraham.ac.uk

Babraham Institute

UNIVERSITY OF CAMBRIDGE

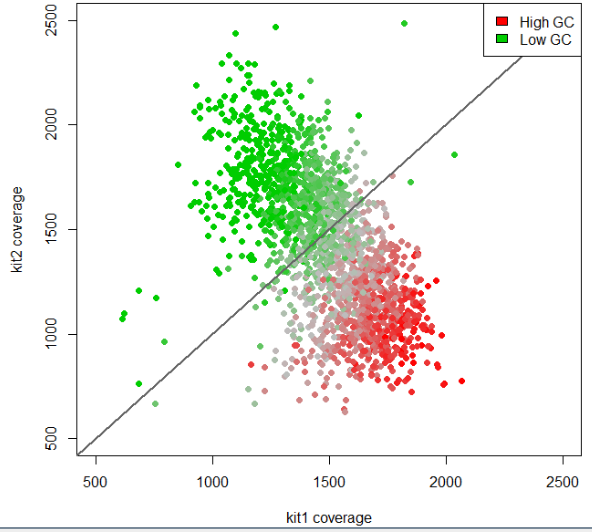
Babraham Bioinformatics

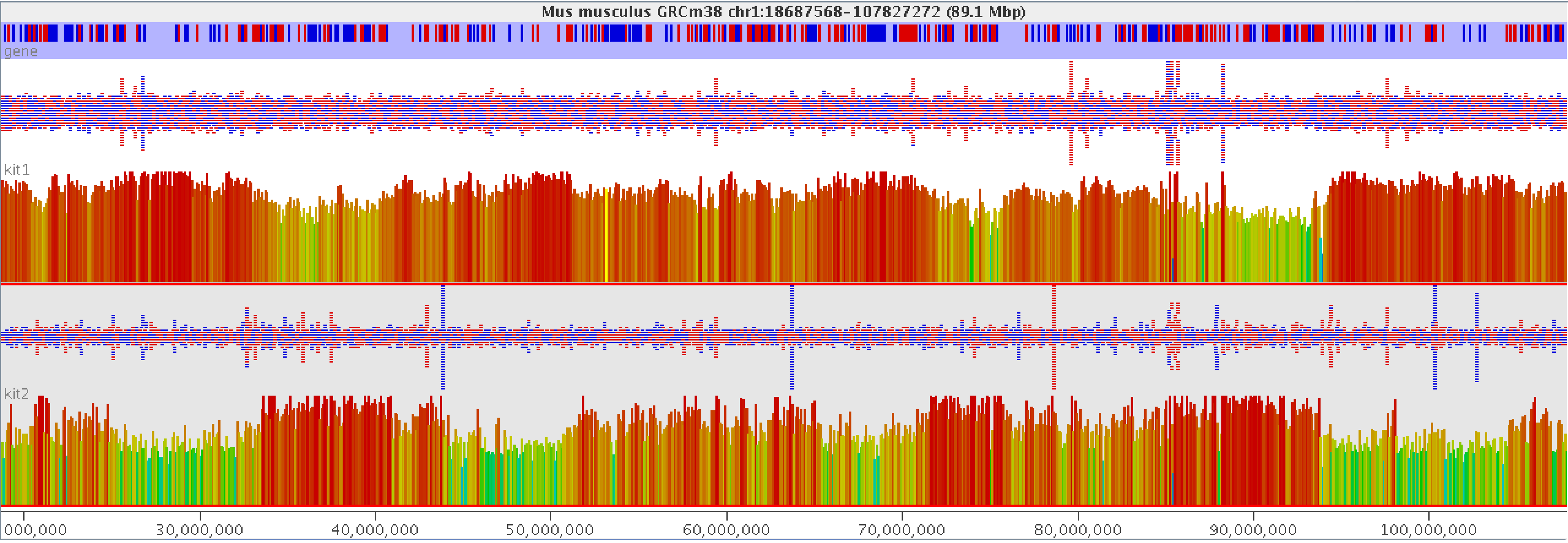
Biases

- All datasets contain biases
  - Technical
  - Biological
  - Statistical
- Biases can lead to incorrect conclusions
- We should be trying to spot these
  - Some are more obvious than others!

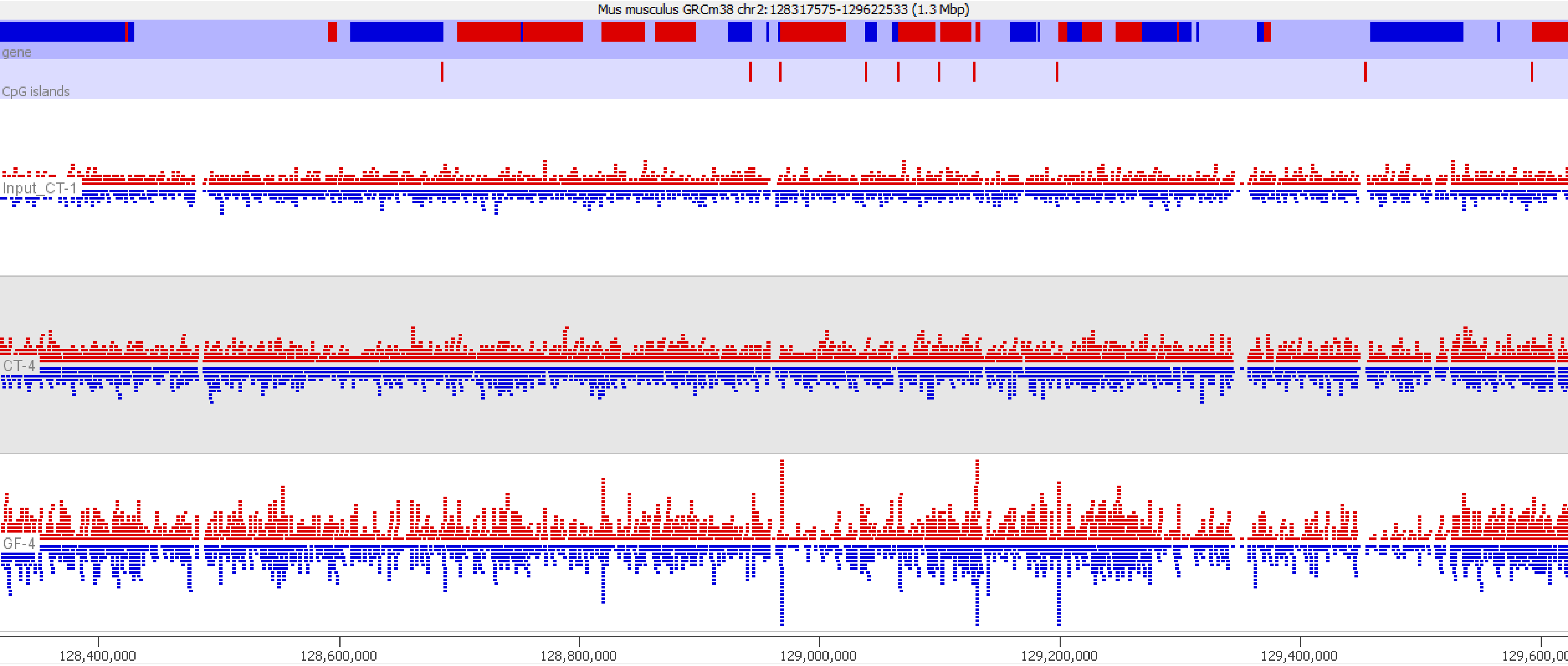
Technical Biases

- Simple GC bias from different polymerases in PCR





Technical Bias



Technical Biases

- Mass Spec Data
  - Membrane proteins underrepresented
    - Hydrophobic
    - Lipid Environment

Statistical Biases

- The power to detect a significant effect is based on:
  - How big the change is
  - How well observed the data is (sample size)

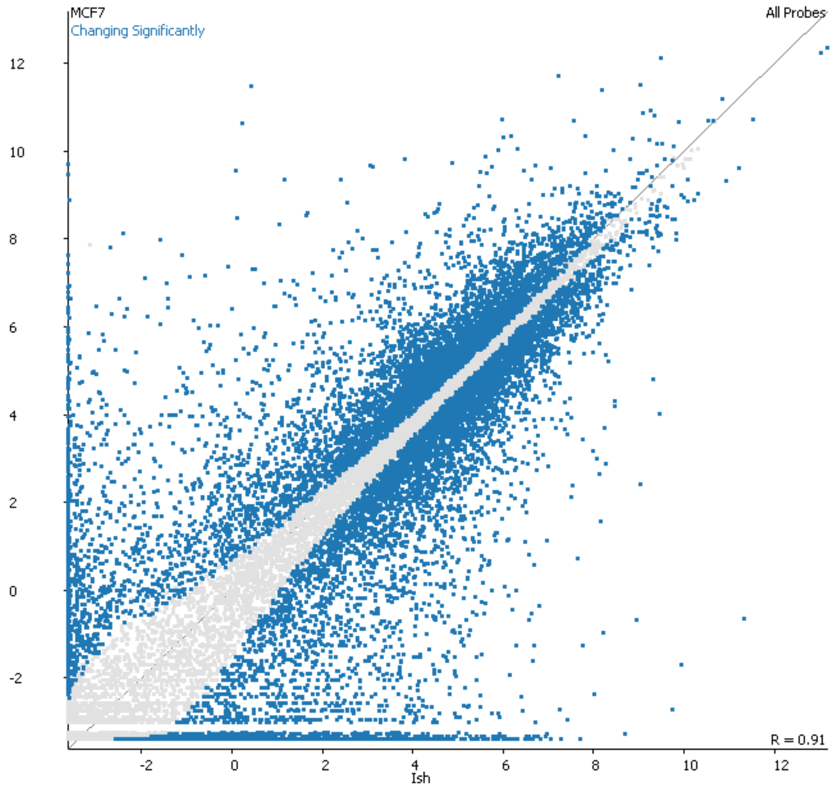
- Lists of hits are often biased based on statistical power

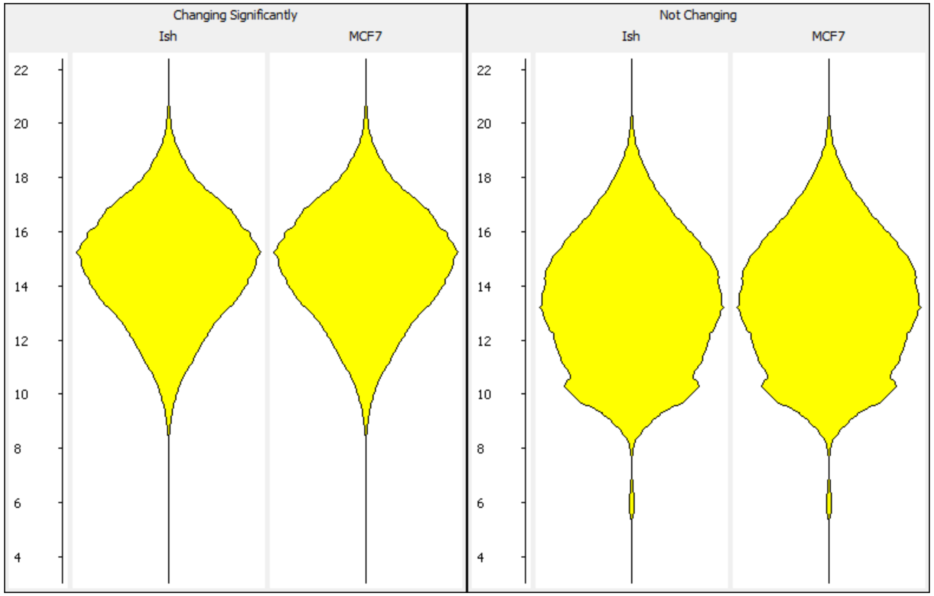
RNA-Seq Statistical Biases

What determines whether a gene is identified as significantly differentially regulated?

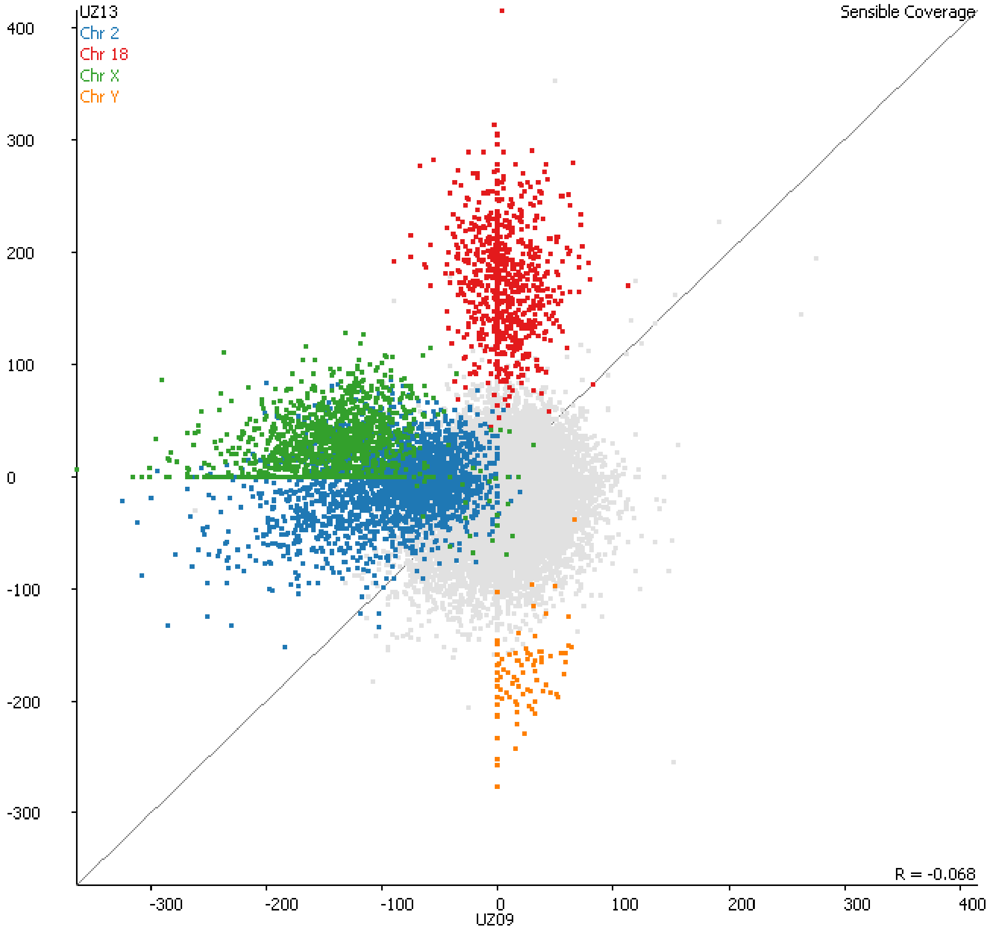
- The amount of change (fold change)
- The variability
- How well observed was it
  - How much sequencing was done overall?
  - How highly expressed was the gene?
  - How long was the gene?
  - How mappable was the gene?

RNA-Seq Statistical Biases





Biological Biases



Biases Look Like Real Biology

Bias	**Function**	**P-Value**
High GC	DNA-Templated Transcription	2.00E-20
Low GC	GPCR Signalling	4.00E-12
Long Genes	Synapse	2.30E-30
Chr 18	Homophilic Cell Adhesion	1.01E-26

Hindawi Publishing Corporation
Evidence-Based Complementary and Alternative Medicine
Volume 2016, Article ID 7276161, 13 pages
http://dx.doi.org/10.1155/2016/7276161

Research Article

Epigenetic Profiling of H3K4Me3 Reveals Herbal Medicine Jinfukang-Induced Epigenetic Alteration Is Involved in Anti-Lung Cancer Activity

Jun Lu,[1] Xiaoli Zhang,[1] Tingting Shen,[1] Chao Ma,[2] Jun Wu,[1] Hualei Kong,[1] Jing Tian,[3] Zhifeng Shao,[1] Xiaodong Zhao,[1,2] and Ling Xu[2,4]

[1] Shanghai Center for Systems Biomedicine, School of Biomedical Engineering, State Key Laboratory on Oncogene and Bio-ID Center, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
[2] Tumor Institute of Traditional Chinese Medicine, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, 725 South Wanping Road, Shanghai 200032, China
[3] College of Life Science, Northwest University, 229 Taibai Road, Xi'an 710069, China

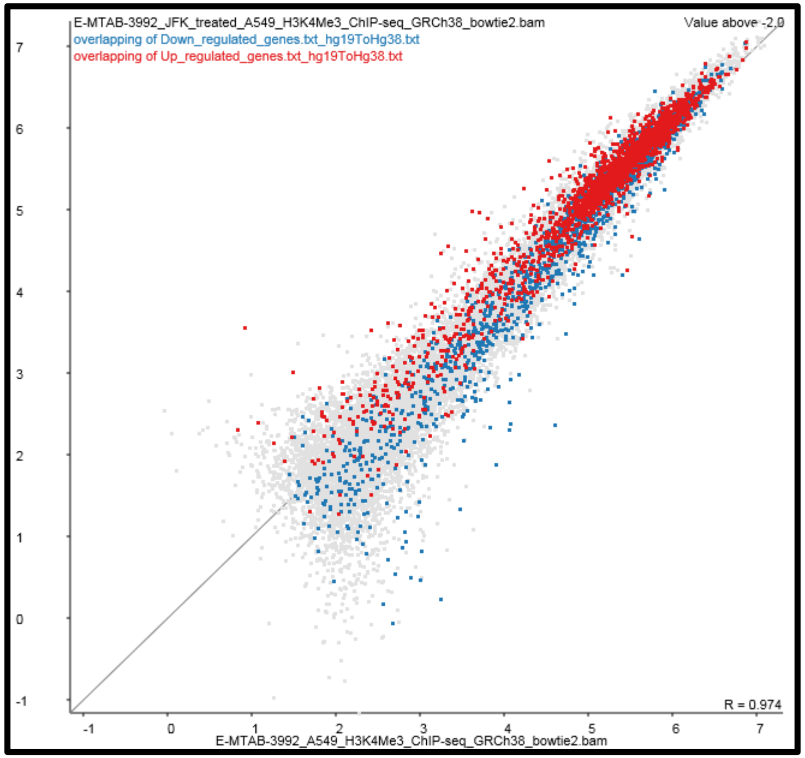
Gene Ontology analysis indicates that these genes are involved in tumor-related pathways, including pathway in cancer, basal cell carcinoma, apoptosis, induction of programmed cell death, regulation of transcription (DNA-templated), intracellular signal transduction, and regulation of peptidase activity.

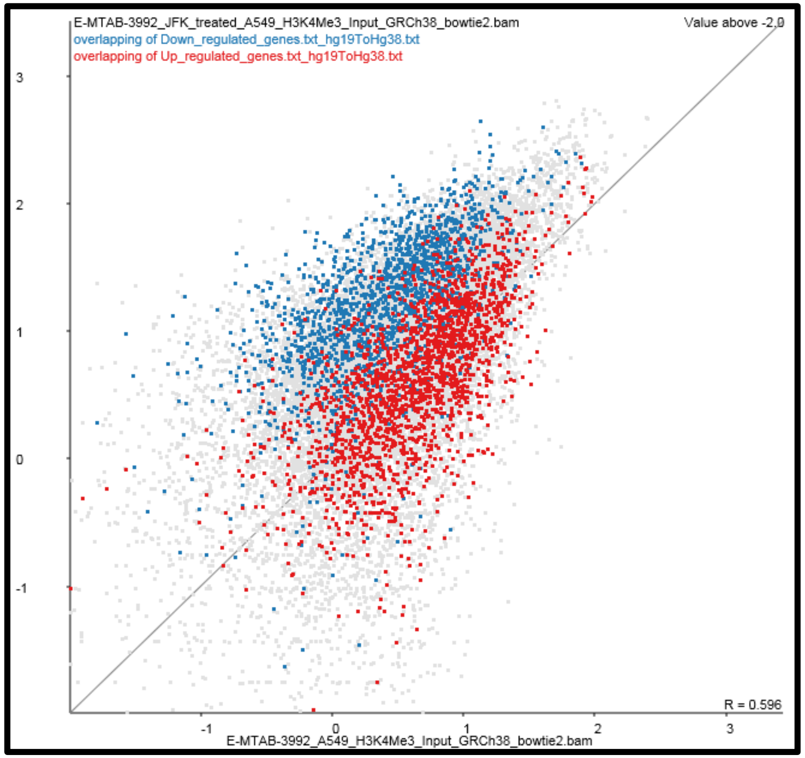
Traditional Chinese medicine Jinfukang (JFK) has been clinically used for treating lung cancer. To examine whether epigenetic modifications are involved in its anticancer activity, we performed a global profiling analysis of H3K4Me3, an epigenomic marker associated with active gene expression, in JFK-treated lung cancer cells. We identified 11,670 genes with significantly altered status of H3K4Me3 modification following JFK treatment ($P < 0.05$). Gene Ontology analysis indicates that these genes are involved in tumor-related pathways, including pathway in cancer, basal cell carcinoma, apoptosis, induction of programmed cell death, regulation of transcription (DNA-templated), intracellular signal transduction, and regulation of peptidase activity. In particular, we found that the levels of H3K4Me3 at the promoters of *SUSD2*, *CCND2*, *BCL2A1*, and *TMEM158* are significantly altered in A549, NCI-H1975, NCI-H1650, and NCI-H2228 cells, when treated with JFK. Collectively, these findings provide the first evidence that the anticancer activity of JFK involves modulation of histone modification at many cancer-related gene loci.

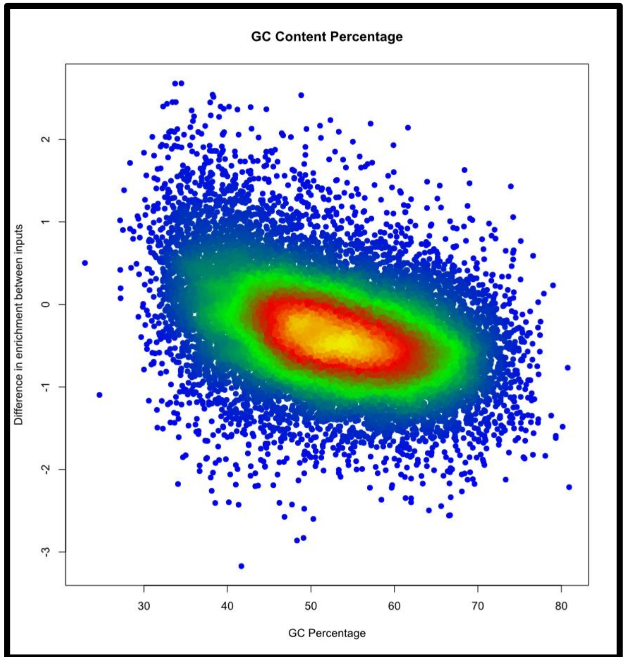
1. Introduction

Chromatin is the macromolecular complex of DNA and histone proteins that provides the scaffold for packaging the eukaryotic genome [1, 2]. Histones H2A, H2B, H3, and H4 are the basic components of nucleosomes, which form the fundamental unit of chromatin [3, 4]. Chemical modifications to the histones alter chromatin structure and regulate gene expression by altering noncovalent interactions within and between nucleosomes [2, 5]. H3K4Me3 is an active histone modification which is positively associated with gene expression [3, 6]. Previous studies have shown that the levels of H3K4Me3 modification are closely associated with the development, treatment, and diagnosis of disease [7–9]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been developed to systematically characterize the contribution of epigenetic regulation in various biological processes via genome-wide profiling of various chemical modifications of histone proteins and genomic DNA methylation [10].

Lung cancer has become the leading cause of cancer-related deaths worldwide [11]. Overall, only 16.8% of patients with lung cancer survive five years after their first definite diagnosis, mainly as a consequence of uncontrollable cell proliferation or tumor metastasis [12, 13]. Although various therapeutic interventions, including surgery, chemotherapy, and radiotherapy, have been developed to prolong the survival time of patients, drug side effects, pain, and emaciation



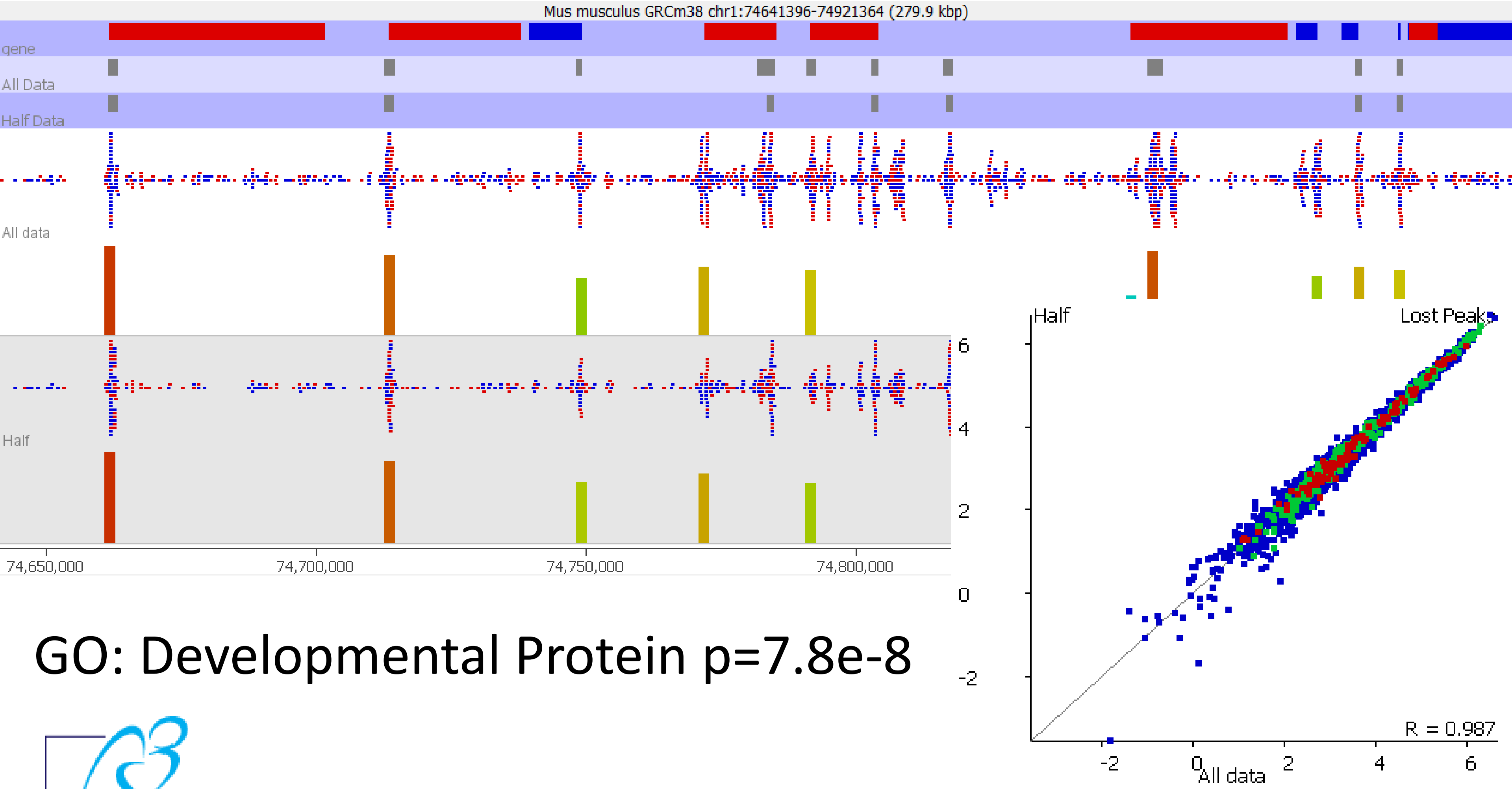




Babraham Institute

Babraham Bioinformatics

Not Significant ≠ Not Changing



GO: Developmental Protein p=7.8e-8

Babraham Institute

UNIVERSITY OF CAMBRIDGE

Babraham Bioinformatics

Relating Hits to Genes

- Most functional analysis is done at the gene level
  - Gene Ontology
  - Pathways
  - Interactions
- Many hits are not gene based
- Power differences can affect this too

Random Genomic Positions

- Find closest gene
  - Synapse, Cell Junction, postsynaptic membrane (p=8.9e-12)
  - Membrane (p=4.3e-13)
  - Glycoprotein (p=1.3e-12)
- Find overlapping genes
  - Plekstrin homology domain (p=1.8e-7)
  - Ion transport (p=7.1e-7)
  - ATP-binding (p=3.8e-8)

Random Transcripts

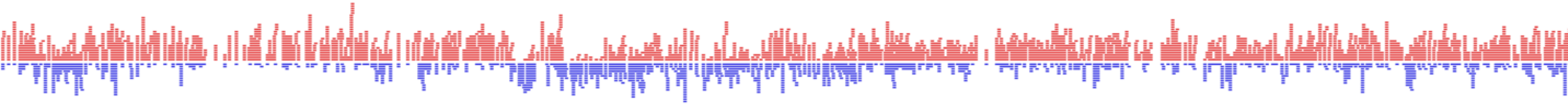
- Tends to favour genes with more splice variants
  - Metal Binding, Zinc Finger ($p=4.4e-12$)
  - Nucleus, Transcription Regulation ($p=2.4e-14$)

Identifying and Correcting Biases

- Address biases during:
  - Planning
  - Quantitation
  - Initial exploration
  - Hit validation
  - GO/Pathway Analysis

Minimising Bias

- Can I adjust my analysis to minimise bias?



- Fixed Size Windows
  - Biased towards higher coverage
  - Biased by CpG density
  - Favours CpG islands
  - Misses many real hits

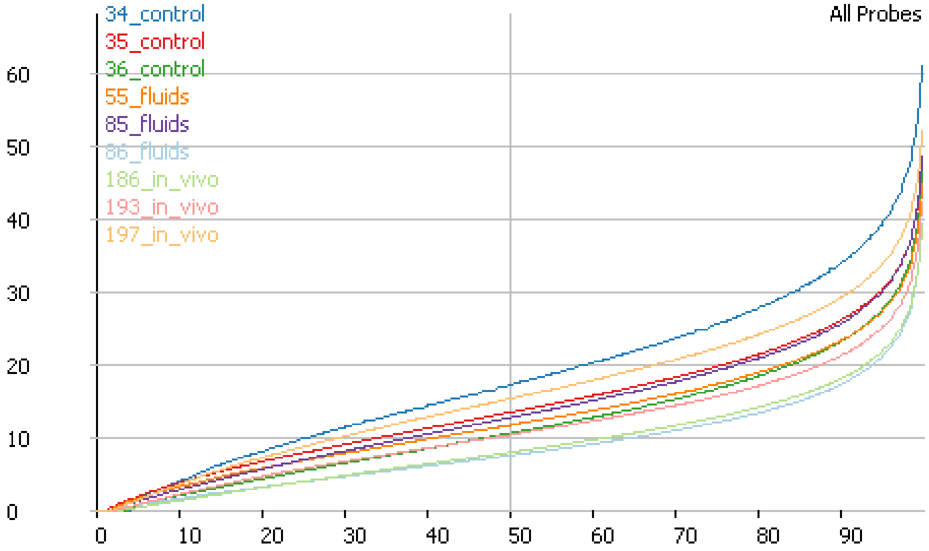
- Fixed Data Windows
  - Even noise profile
  - Even statistical power
  - Uneven resolution
  - Easier to interpret
  - Misses fewer hits

Normalising / Regression

- Maths can fix my data!

- Understand first
- Minimal correction

- Loss of interpretability

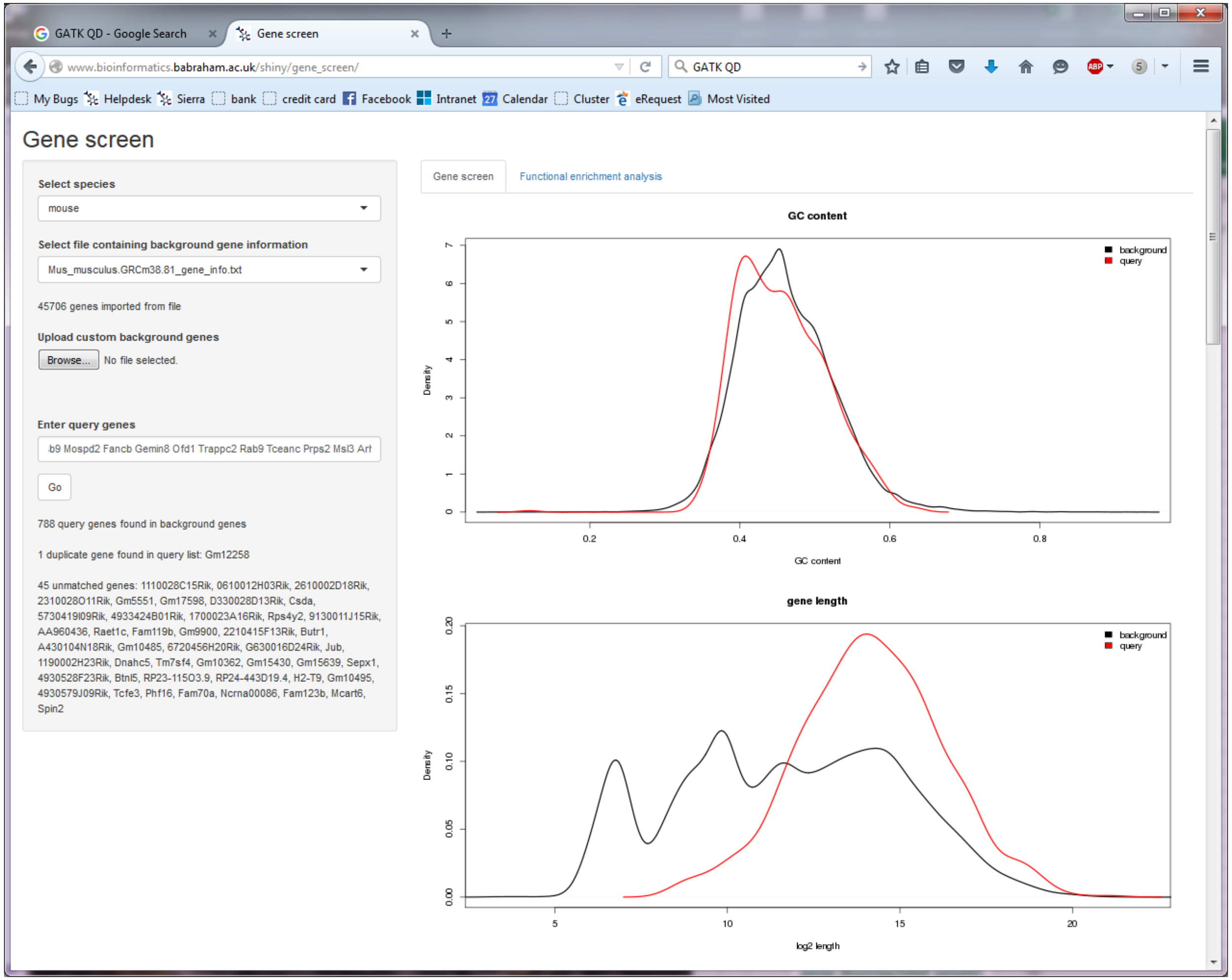


Hit Validation

- Do my hits look different from non-hits in factors which should be unrelated

- How easy would it be for the effect I see to be generated through a technical artefact?

Look for confounders



Gene screen

Select species: mouse

Select file containing background gene information: Mus_musculus.GRCm38.81_gene_info.txt

45706 genes imported from file

Upload custom background genes: Browse... No file selected.

Enter query genes: b9 Mospd2 Fancb Gemin8 Ofd1 Trappc2 Rab9 Tceanc Prps2 Msl3 Arl

Go

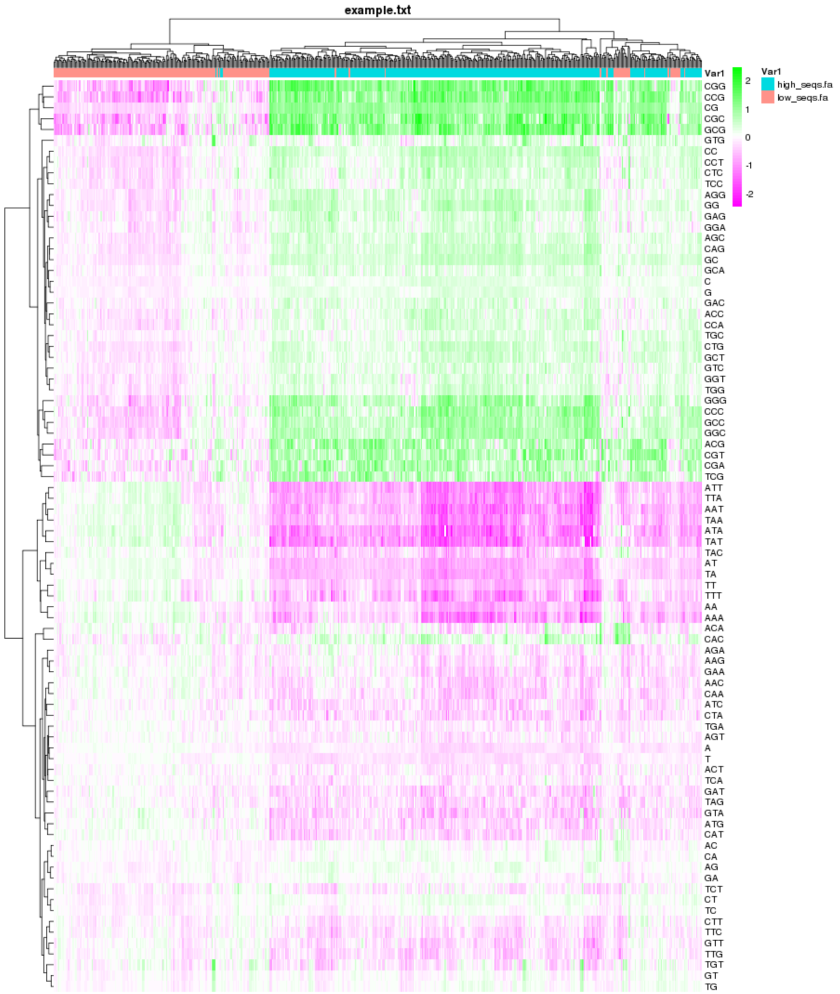
788 query genes found in background genes

1 duplicate gene found in query list: Gm12258

45 unmatched genes: 1110028C15Rik, 0610012H03Rik, 2610002D18Rik, 2310028O11Rik, Gm5551, Gm17598, D330028D13Rik, Csda, 5730419I09Rik, 4933424B01Rik, 1700023A16Rik, Rps4y2, 9130011J15Rik, AA960436, Raet1c, Fam119b, Gm9900, 2210415F13Rik, Butr1, A430104N18Rik, Gm10485, 6720456H20Rik, G630016D24Rik, Jub, 1190002H23Rik, Dnahc5, Tm7sf4, Gm10362, Gm15430, Gm15639, Sepx1, 4930528F23Rik, Btnl5, RP23-115O3.9, RP24-443D19.4, H2-T9, Gm10495, 4930579J09Rik, Tcfe3, Phf16, Fam70a, Ncrna00086, Fam123b, Mcart6, Spin2

Look for confounders

- Compter
  - Sequence kmer analysis
  - Does composition explain my hits?

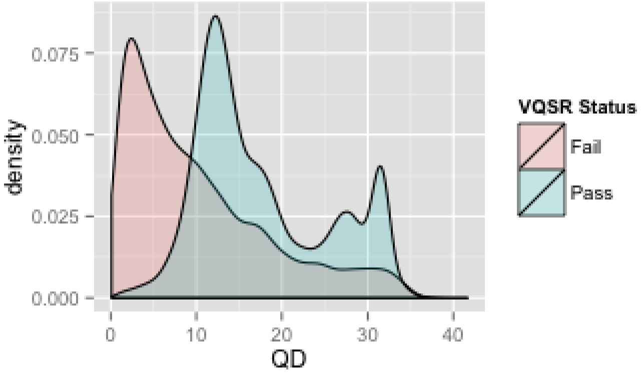


www.bioinformatics.babraham.ac.uk/projects/compter

SNP Filtering

- Factors to consider
  - Depth of coverage
  - Base call Quality
  - Mapping Quality
  - Position within read
  - Strand bias
  - Genomic Position

- Types of problem
  - Too low OR high
  - Different to non-hits
    - Error
    - Bias
    - Biology



GO / Pathway Analysis

- Make sure you're asking the right question

- Background models are key
  - What is different between hits and genome

  - What is different between actual hits and possible hits