Artefacts and Biases in Gene Set Analysis

Simon Andrews, Laura Biggins, Christel Krueger

simon.andrews@babraham.ac.uk
laura.biggins@babraham.ac.uk
christel.krueger@babraham.ac.uk

v2017-06
What does gene set enrichment test?

• Is a functional gene set enriched for genes in my hit list compared to a background set

• Are some genes more likely to turn up in the hits for technical reasons?

• Are some genes never likely to turn up in the hit list for technical reasons?
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
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
Not Significant ≠ Not Changing

GO: Developmental Protein  p=7.8e-8
<table>
<thead>
<tr>
<th>Bias</th>
<th>Function</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High GC</td>
<td>DNA-Templated Transcription</td>
<td>2.00E-20</td>
</tr>
<tr>
<td>Low GC</td>
<td>GPCR Signalling</td>
<td>4.00E-12</td>
</tr>
<tr>
<td>Long Genes</td>
<td>Synapse</td>
<td>2.30E-30</td>
</tr>
<tr>
<td>Chr 18</td>
<td>Homophilic Cell Adhesion</td>
<td>1.01E-26</td>
</tr>
</tbody>
</table>
Research Article

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

Jun Lu, Xiaoli Zhang, Tingting Shen, Chao Ma, Jun Wu, Hualei Kong, Jing Tian, Zhifeng Shao, Xiaodong Zhao, and Ling Xu

1Shanghai 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
2Tumor Institute of Traditional Chinese Medicine, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, 725 South Wanning Road, Shanghai 200032, China
3College 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 1,670 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, BCL2AI, and TIMM13 are significantly altered in A549, NCI-H460, and NCI-H1228 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, 7]. 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 emotion...
Correct selection of a background list can make a huge difference

• What genes were you likely to see?
  – Some are technically impossible
    • Membrane proteins in LC-MS
    • Small-RNA in RNA-Seq
  – Some are much less likely
    • Unexpressed or low expressed in RNA-Seq
    • Unmappable in ChIP-Seq
    • Low CpG content in BS-Seq
Statistical biases affect gene sets too

- Fisher’s test is powered by
  - Magnitude of change
  - Observation level

- Big lists have more power to detect change
- Small lists are very difficult to detect
Relating Hits to Genes

• Most functional analysis is done at the gene level
  – Gene Ontology
  – Pathways
  – Interactions
• Many hits are not gene based
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)
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

- Make sure background is appropriate
- Check for compositional bias
- Be suspicious of some ontology categories: ribosomal, cytoskeleton, extracellular, secreted, translation
Look for confounders

http://www.bioinformatics.babraham.ac.uk/shiny/gene_screen/
Look for confounders

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