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

V2020-10

Babraham Bioinformatics

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?

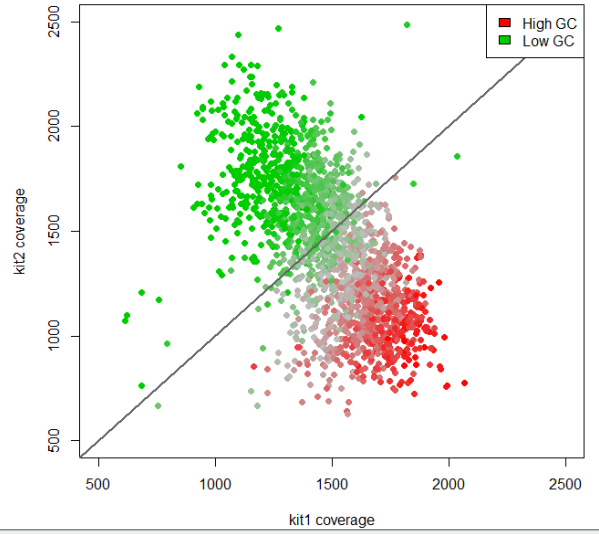
- Are some genes **never likely** to turn up in the hit list for technical reasons?

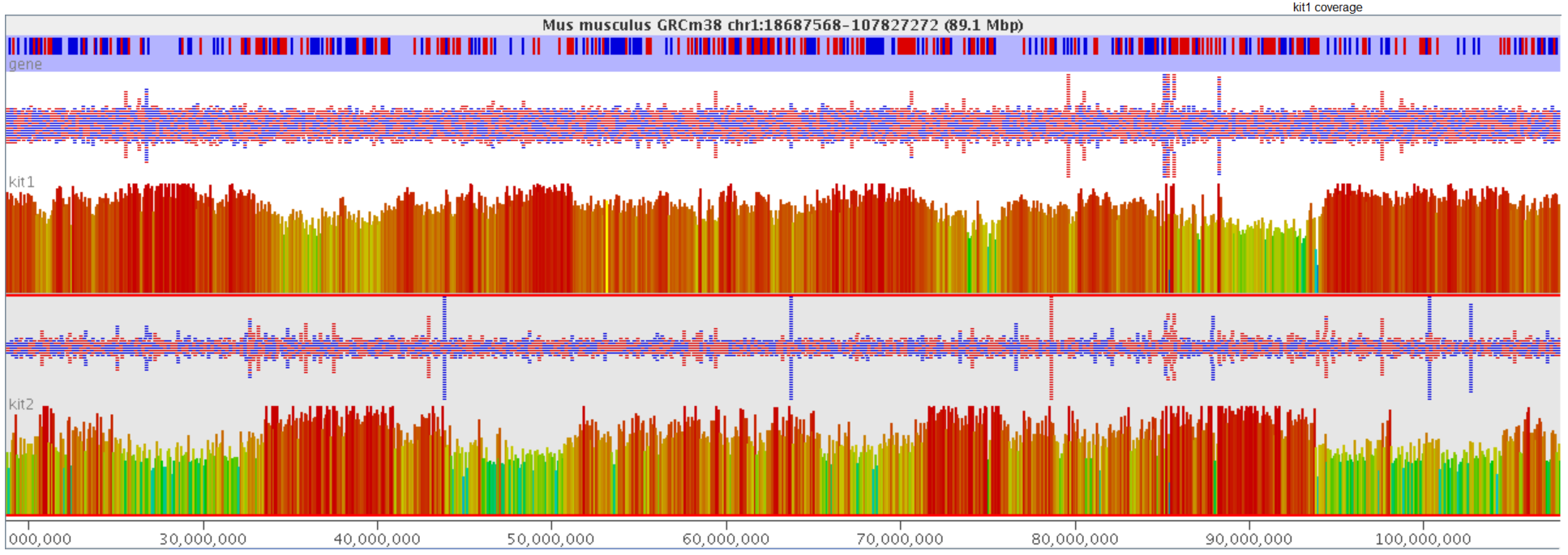
Biases

- 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)

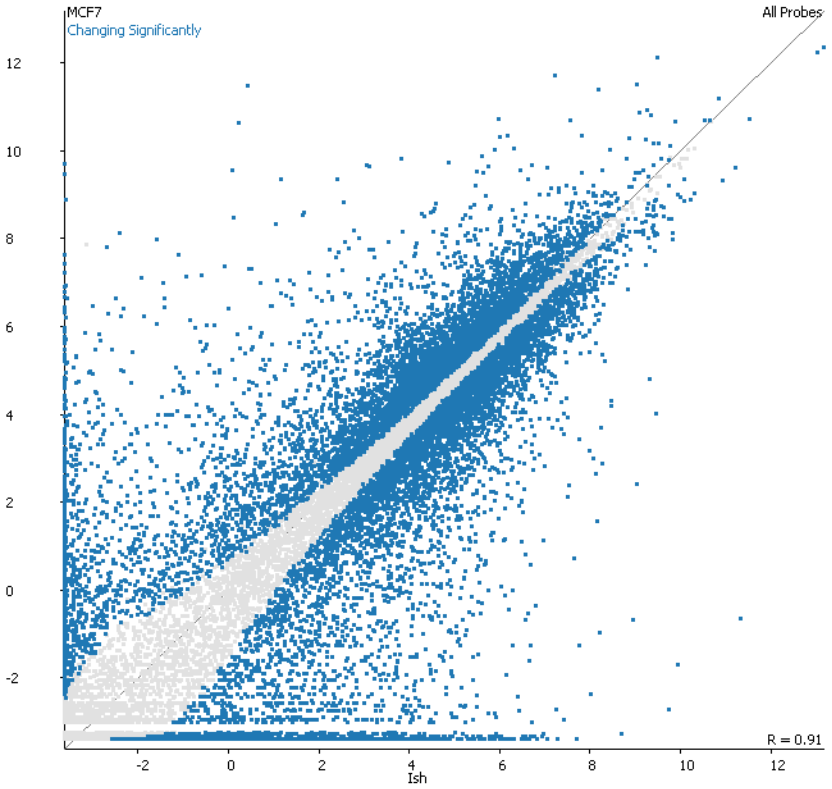
- Lists of hits are often biased based on statistical power

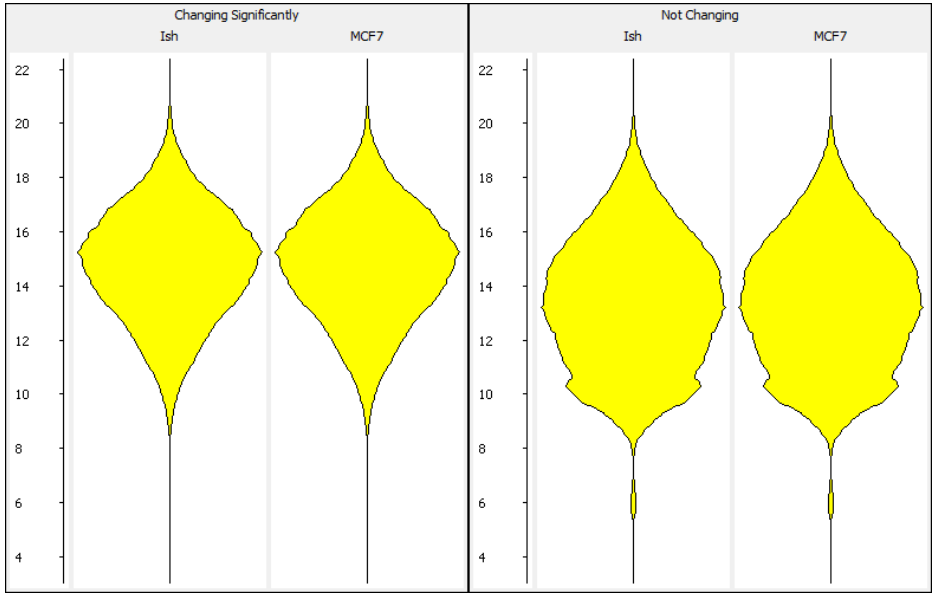
RNA-Seq Statistical Biases

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?

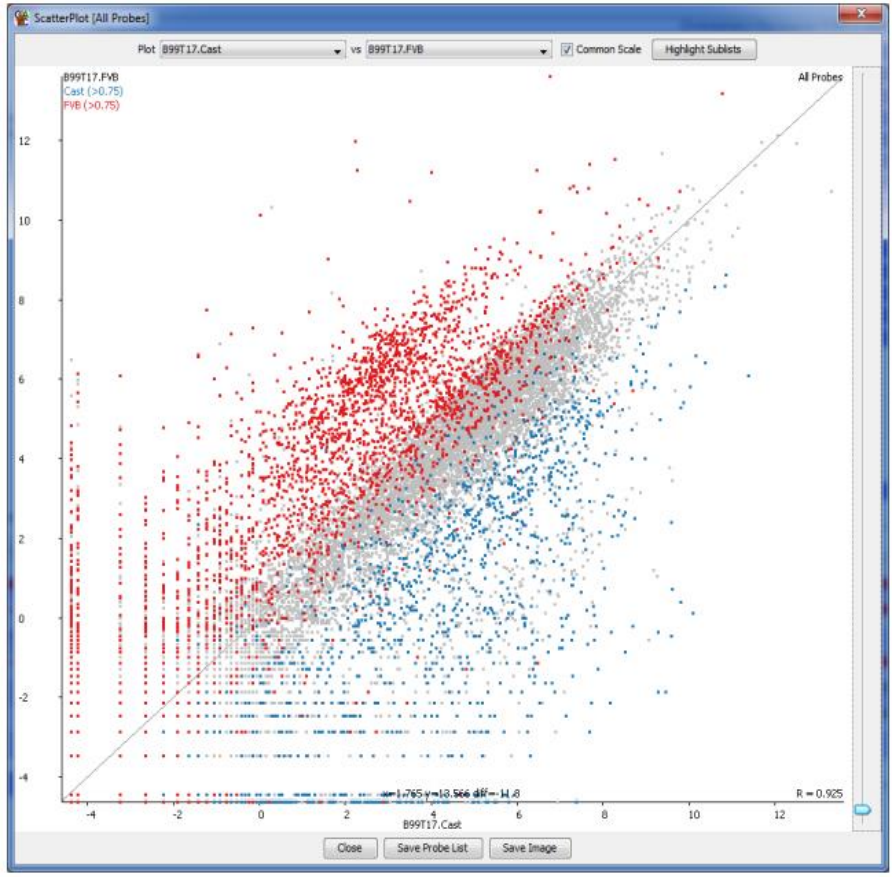
RNA-Seq Statistical Biases

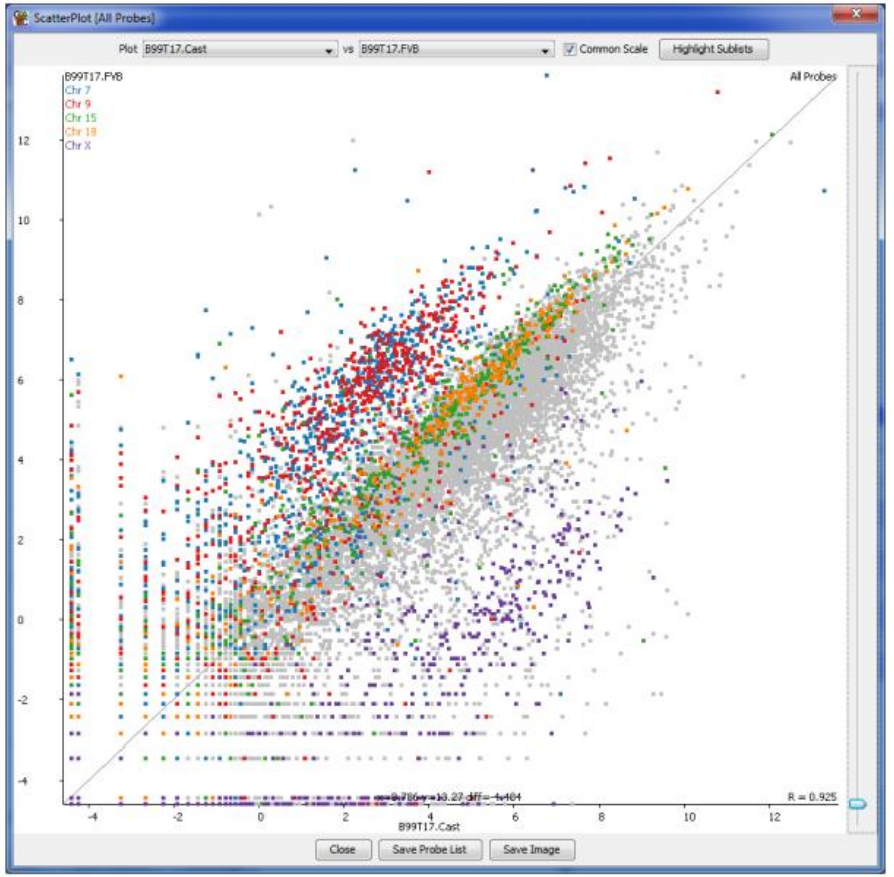




- Unlikely to ever see hits from genes which are
  - Lowly expressed
  - Short

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*

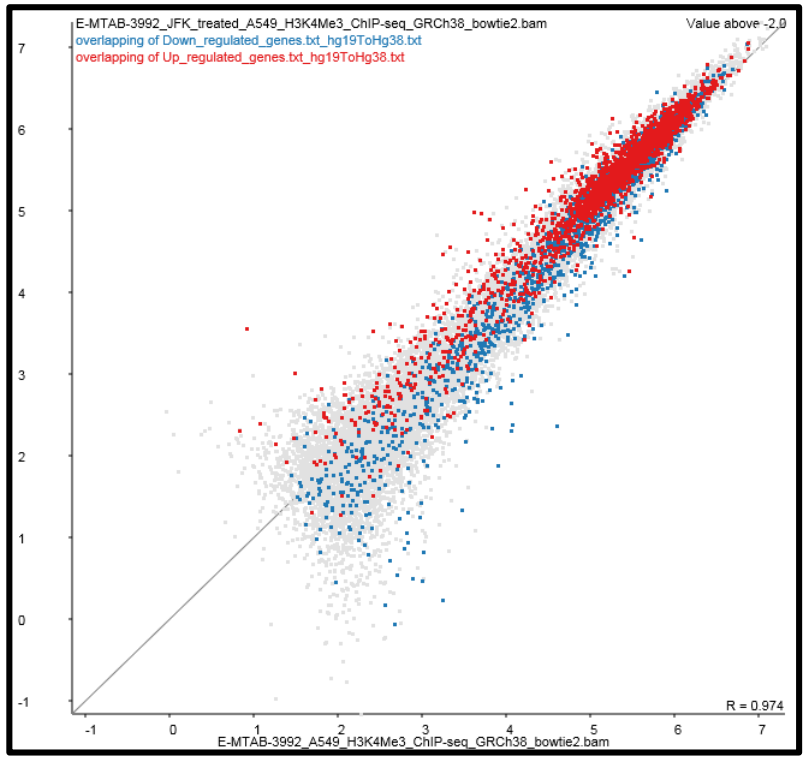
Gene Ontology analysis indicates that these genes are involved in tumor-related pathways, including pathway in cancer, basal cell carcinoma, apoptosis, induction of programmed cell death, regulation of transcription (DNA-templated), intracellular signal transduction, and regulation of peptidase activity.

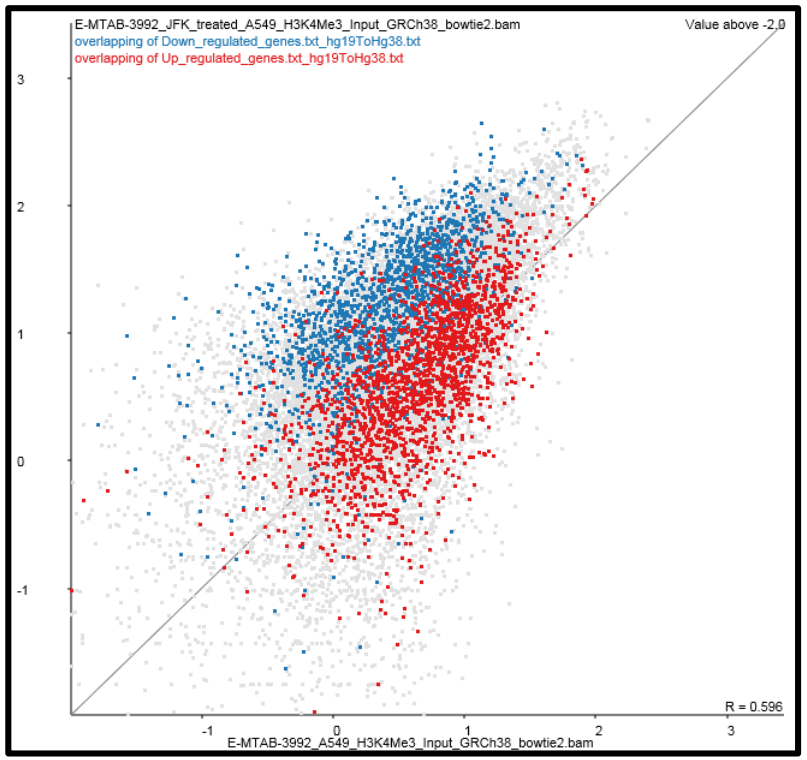
Traditional Chinese medicine Jinfukang (JFK) has been clinically used for treating lung cancer. To examine whether epigenetic modifications are involved in its anticancer activity, we performed a global profiling analysis of H3K4Me3, an epigenomic marker associated with active gene expression, in JFK-treated 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.

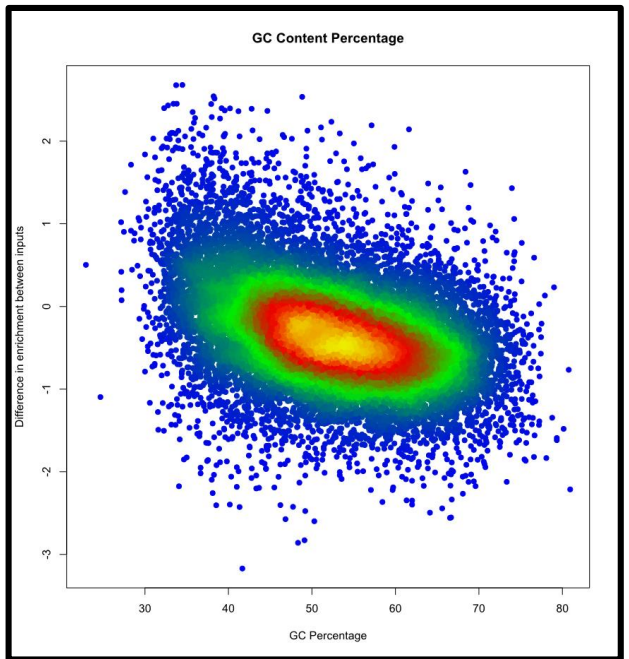
1. Introduction

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


E-MTAB-3992_JFK_treated_A549_H3K4Me3_ChIP-seq_GRCh38_bowtie2.bam
Value above -2.0
overlapping of Down_regulated_genes.txt_hg19ToHg38.txt
overlapping of Up_regulated_genes.txt_hg19ToHg38.txt
R = 0.974
E-MTAB-3992_A549_H3K4Me3_ChIP-seq_GRCh38_bowtie2.bam


E-MTAB-3992_JFK_treated_A549_H3K4Me3_Input_GRCh38_bowtie2.bam
Value above -2.0
overlapping of Down_regulated_genes.txt_hg19ToHg38.txt
overlapping of Up_regulated_genes.txt_hg19ToHg38.txt
R = 0.596
E-MTAB-3992_A549_H3K4Me3_Input_GRCh38_bowtie2.bam


GC Content Percentage
Difference in enrichment between inputs
GC Percentage

What can you do?

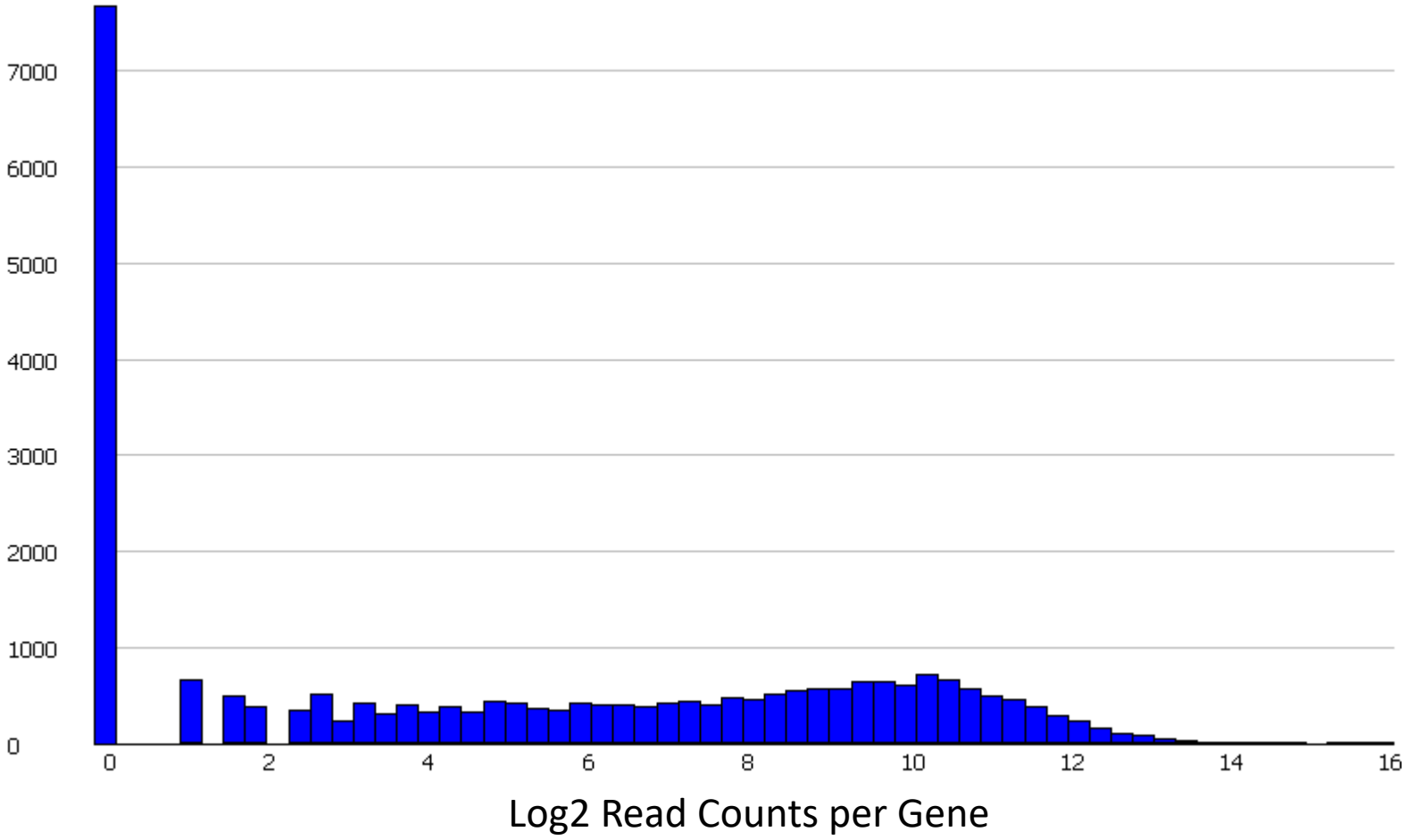
- Think about whether you're likely to have expected biases in your experiment.
  - If possible, restructure to avoid the bias
- Look for unexpected biases.
  - Sometimes the bias ***is*** the interesting biology
- Use custom backgrounds during Gene Set Analysis to help minimise bias (if a tool supports it)

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

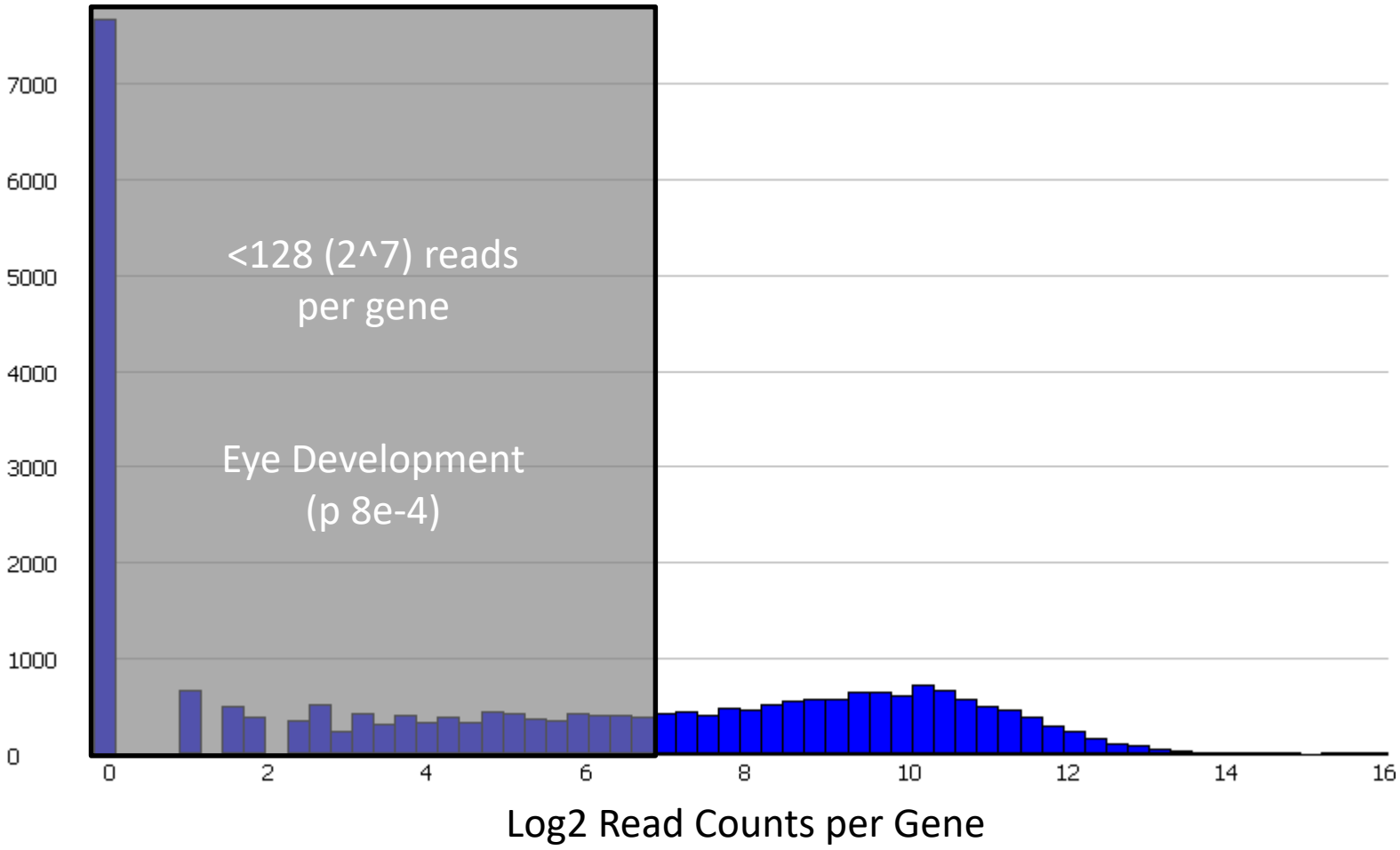
- Make a list of what you ***could*** have seen, and set that as the background.

Expressed Genes



26,127 Genes Measured

Expressed Genes



10,378 Genes Realistically Measured

Statistical biases affect gene sets too

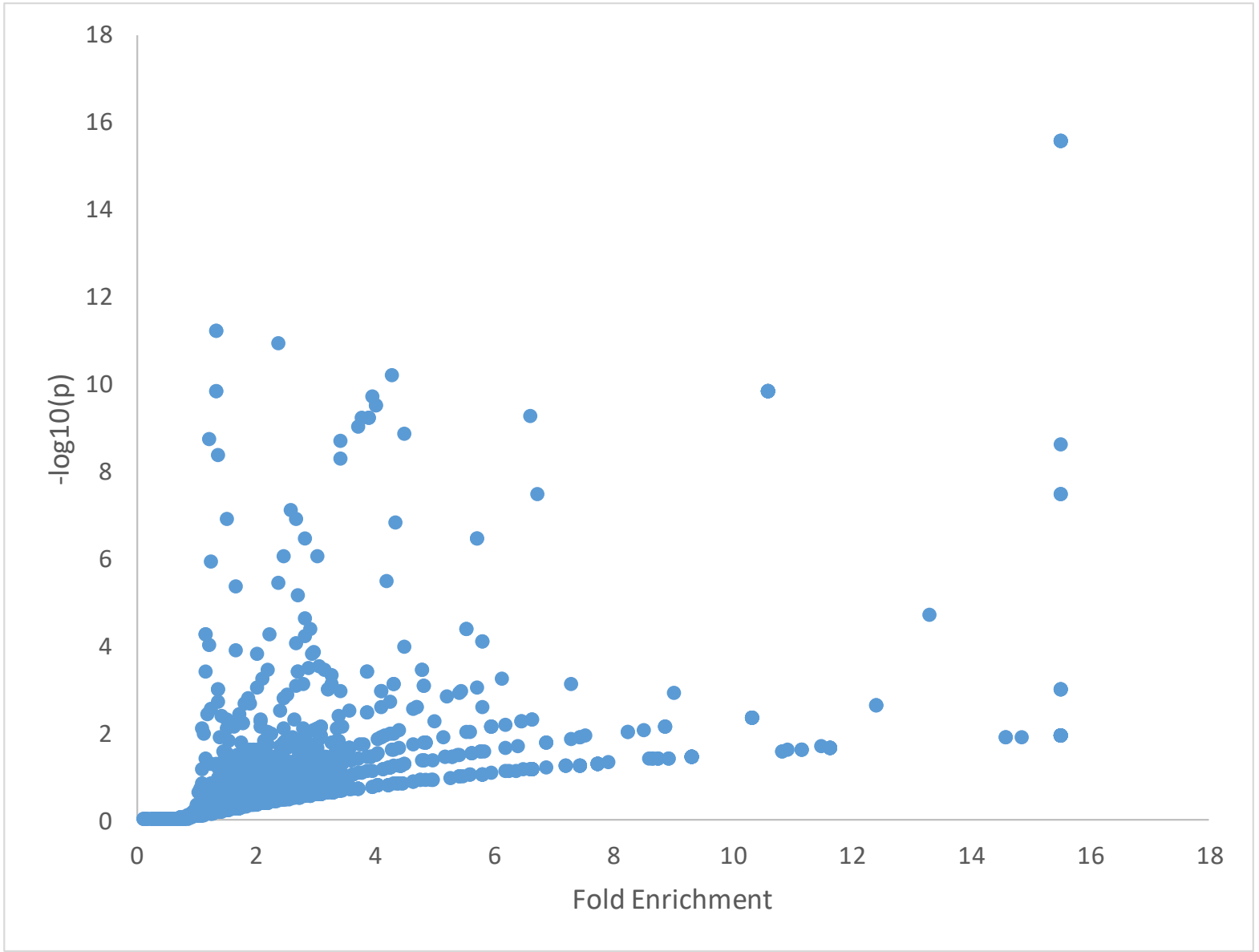
- Fisher's test is powered by
  - Magnitude of change
  - Observation level

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

- Some tools allow you to exclude the largest gene set categories. We often use categories with between 50 – 500 genes in to get power and specificity

- Always look at the enrichment and the p-value when deciding what is interesting

Fold Change and p-value



Other biases: Relating Hits to Genes

- Most functional analysis is done at the gene level
  - Gene Ontology
  - Pathways
  - Interactions

- Many hits are not gene based

Other biases: 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)

Other biases: Random Transcripts

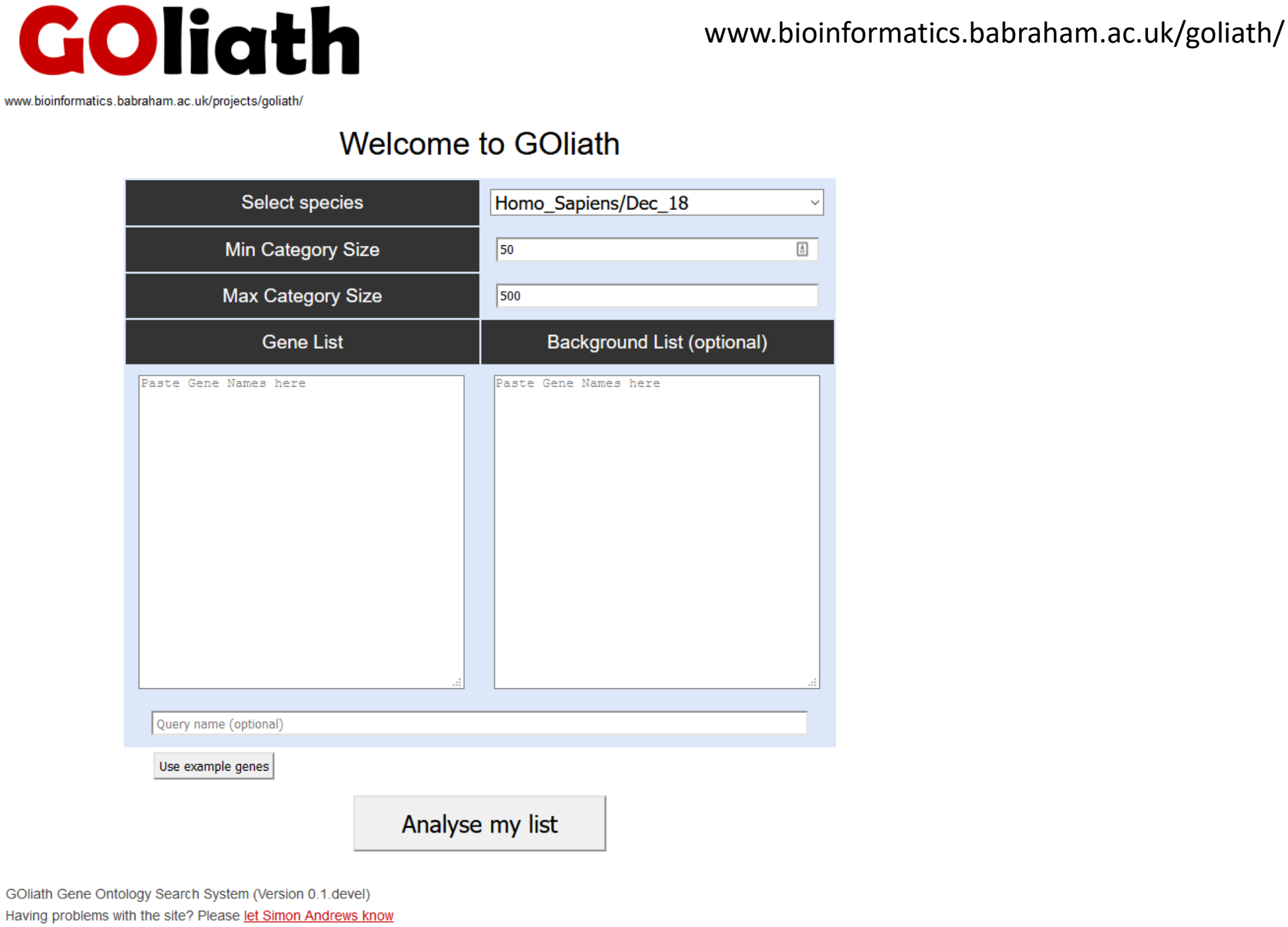
- Tends to favour genes with more splice variants
  - Metal Binding, Zinc Finger (p=4.4e-12)
  - Nucleus, Transcription Regulation (p=2.4e-14)

Stuff which turns up more than it should…

- Did a trawl through GEO RNA-Seq datasets
  - Downloaded pairs of samples which are supposed to be biological replicates
  - Found changing genes
  - Ran GO searches

- Many gene sets give hits. Some categories turn up very often
  - Ribosomal
  - Cytoskeleton
  - Extracellular
  - Secreted
  - Translation

www.bioinformatics.babraham.ac.uk/goliath/



www.bioinformatics.babraham.ac.uk/goliath/

GOliath

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

Results for job Course test

Results Table | Properties | Biases

Hit table

Copy CSV Excel Print

Gene Set	Source	Query count	Background count	Category size	FDR	Enrichment	Potential bias
HALLMARK TNFA SIGNALING VIA NFKB	MSIGDB C2 HALLMARK TNFA SIGNALING VIA NFKB	17	200	200	1.384e-07	9.174	public_data
SIGNALING BY INTERLEUKINS	REACTOME R-HSA-449147.11	20	461	461	4.195e-05	4.683	high_transcripts
POSITIVE REGULATION OF CYTOKINE PRODUCTION	GOBP GO:0001819	16	355	355	0.0005164	4.865	public_data
HALLMARK IL2 STAT5 SIGNALING	MSIGDB C2 HALLMARK IL2 STAT5 SIGNALING	12	200	200	0.0009546	6.476	public_data
HALLMARK APOPTOSIS	MSIGDB C2 HALLMARK APOPTOSIS	10	160	160	0.003579	6.746	
APOPTOSIS	WIKIPATHWAYS 20190910 WP254 HOMO SAPIENS	8	87	87	0.003579	9.925	
REGULATION OF CYTOKINE SECRETION	GOBP GO:0050707	10	154	154	0.003579	7.009	
REGULATION OF INTERLEUKIN-6 PRODUCTION	GOBP GO:0032675	8	101	101	0.007389	8.549	
HALLMARK INFLAMMATORY RESPONSE	MSIGDB C2 HALLMARK INFLAMMATORY RESPONSE	10	200	200	0.008619	5.397	public_data
HALLMARK ALLOGRAFT REJECTION	MSIGDB C2 HALLMARK ALLOGRAFT REJECTION	10	200	200	0.008619	5.397	public_data
Search gene set	Search source	min query / max query	min bg / max bg	min size / max size	min FDR / max FDR	min enrichment / max enrichment	Search bias

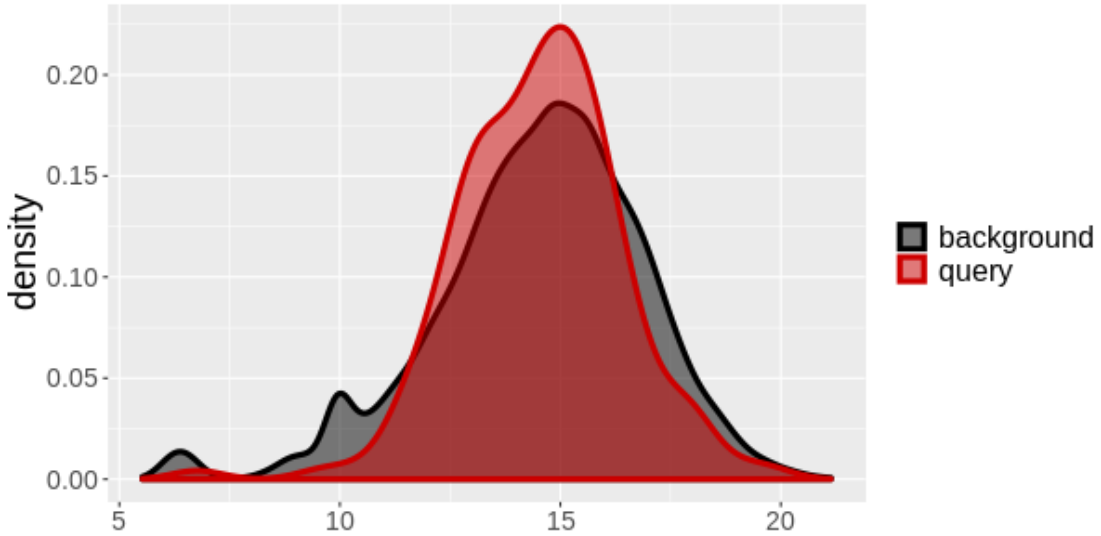
Hit Validation

- Do my hits look different from non-hits in factors which should be unrelated
  - Sequence composition
  - Genomic position
  - Gene Length
  - Number of splice variants
  - etc

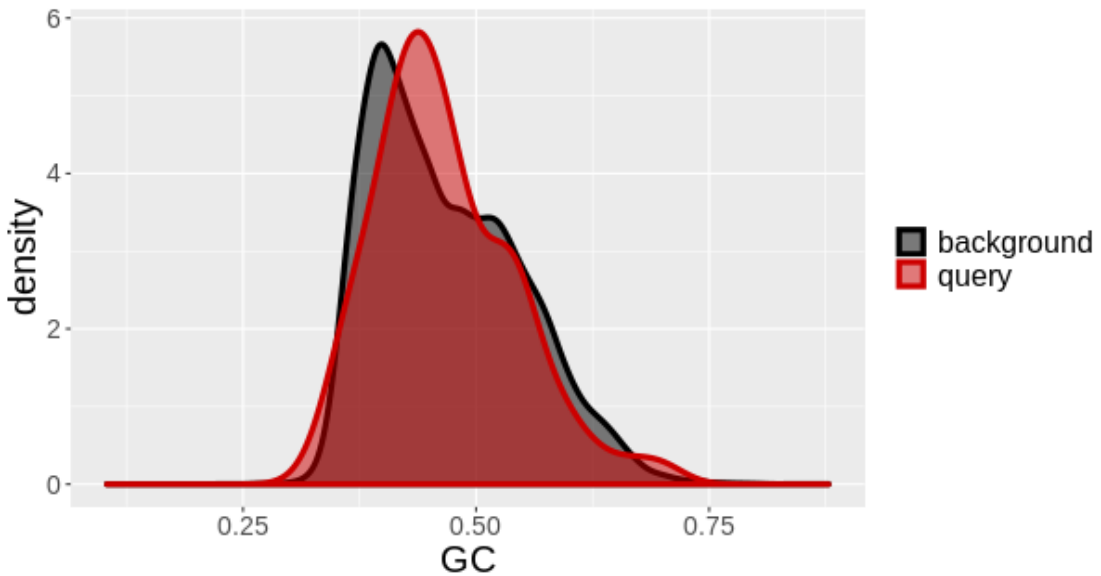
- If a bias exists then is this the actual link between genes? If not then can I fix this by improving my background list?

www.bioinformatics.babraham.ac.uk/goliath/

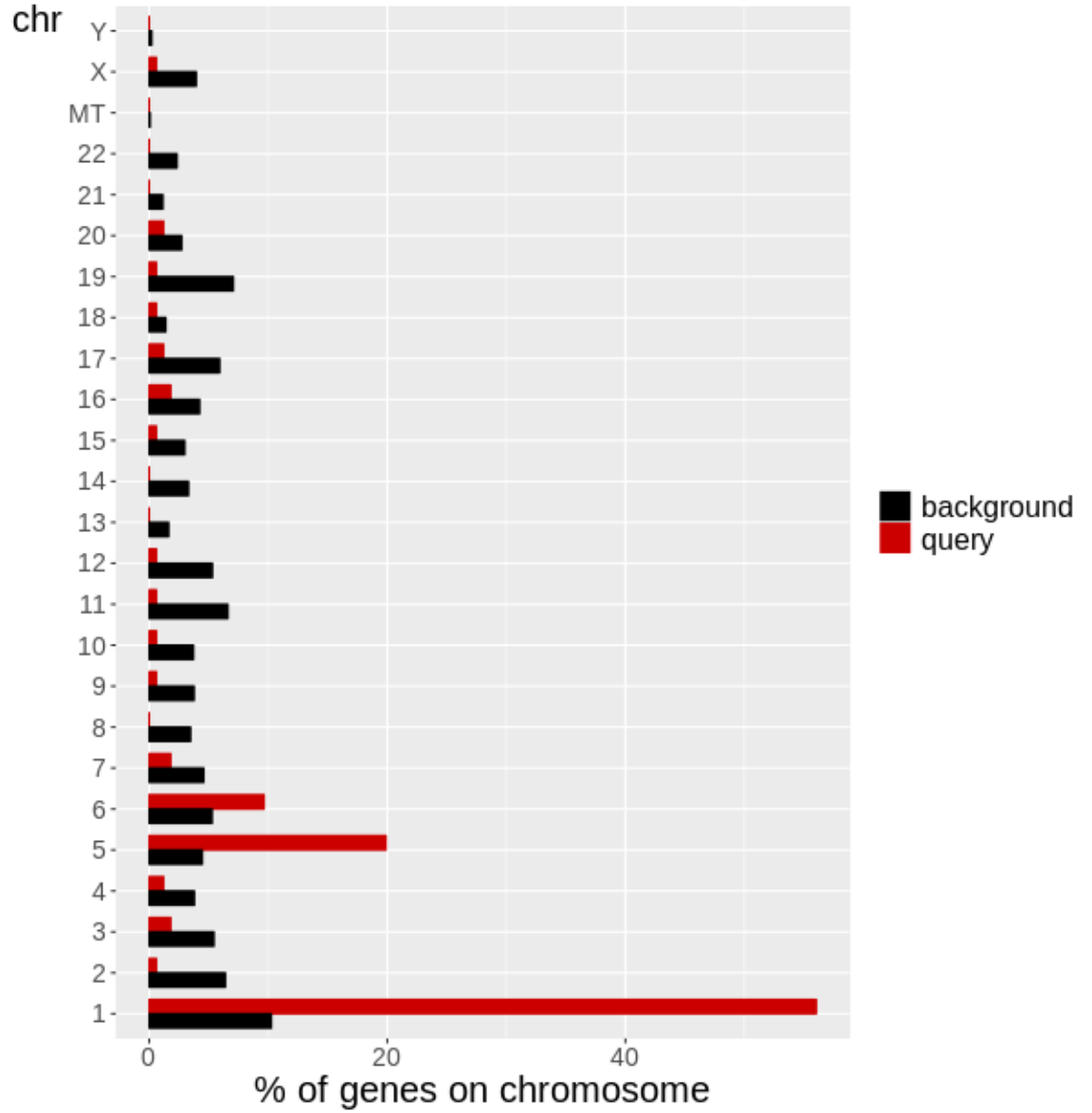
Gene lengths



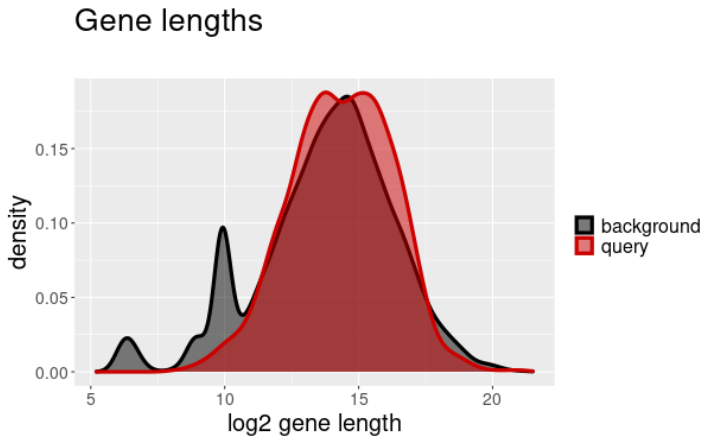
GC content of genes

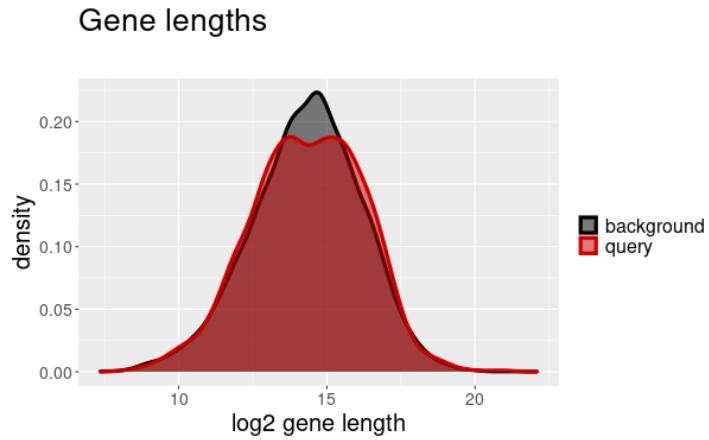


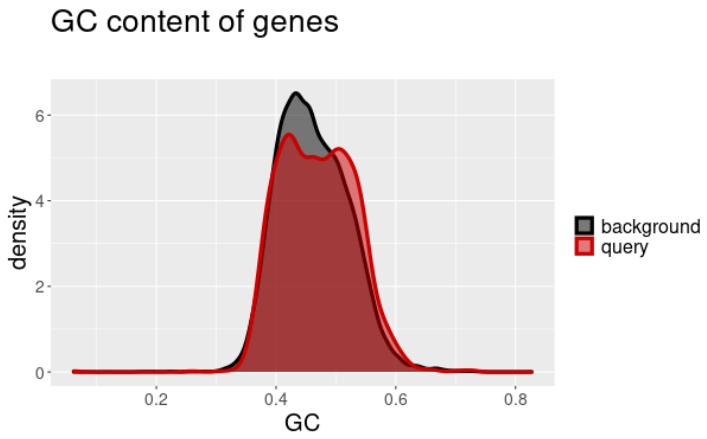
Chromosomal locations

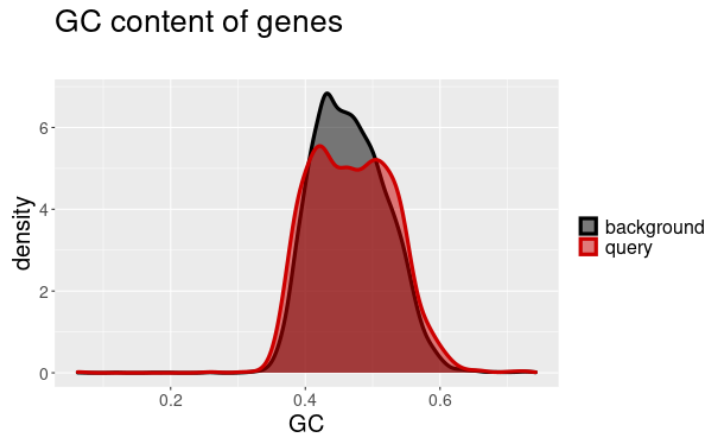


Custom backgrounds can make a difference









Custom backgrounds can make a difference

Top hits without correction

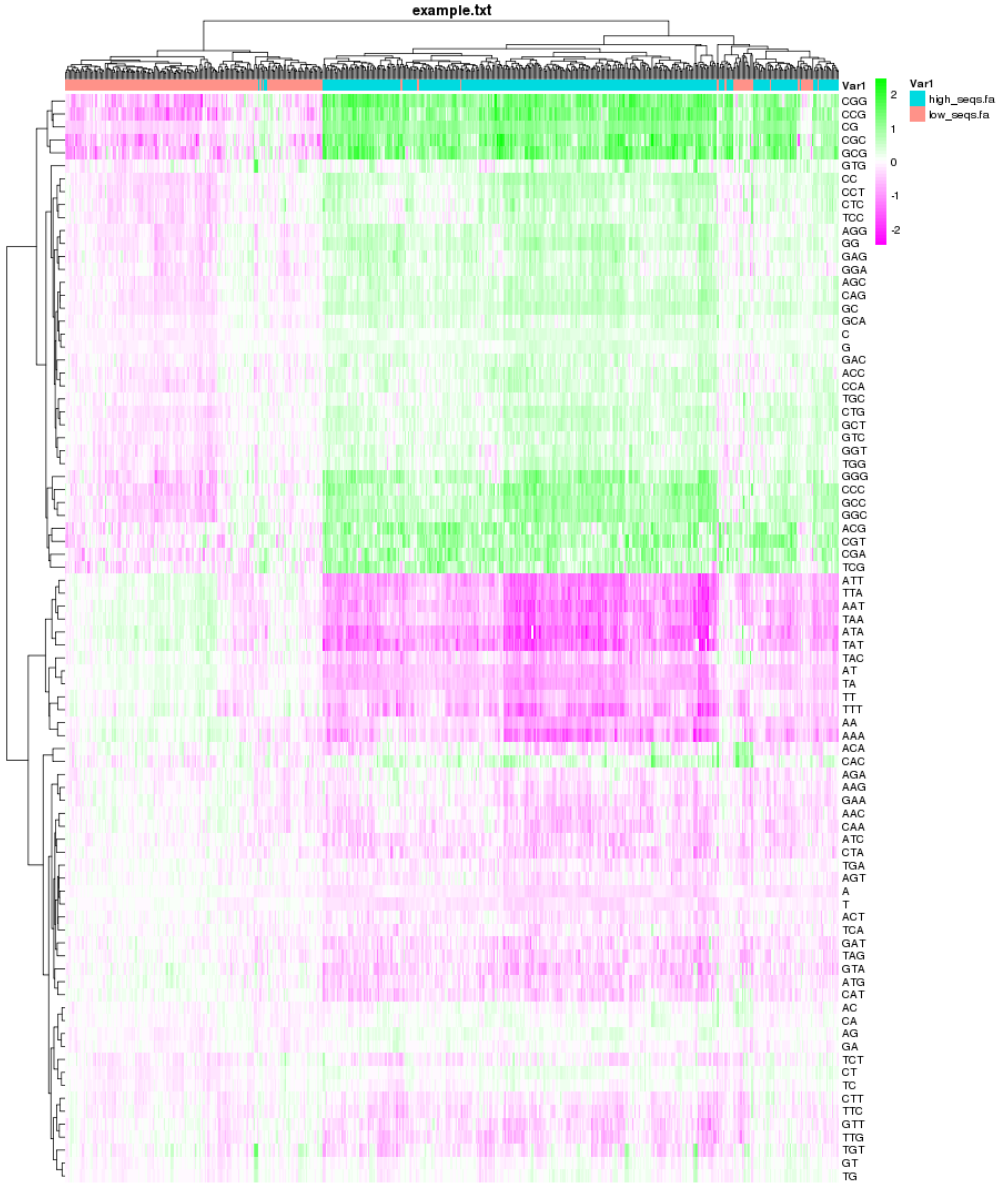
PLURINETWORK
POSITIVE REGULATION OF VASCULATURE DEVELOPMENT
POSITIVE REGULATION OF ANGIOGENESIS
HALLMARK E2F TARGETS
CHROMOSOME, CENTROMERIC REGION
DNA REPAIR
NEGATIVE REGULATION OF CELLULAR AMIDE METABOLISM
POSITIVE REGULATION OF ENDOTHELIAL CELL MIGRATION
NUCLEAR CHROMOSOME SEGREGATION
PID INTEGRIN1 PATHWAY

Top hits with correction

POSITIVE REGULATION OF VASCULATURE DEVELOPMENT
POSITIVE REGULATION OF ANGIOGENESIS
PID INTEGRIN1 PATHWAY
BETA1 INTEGRIN CELL SURFACE INTERACTIONS
INTEGRIN BINDING
ASSEMBLY OF COLLAGEN FIBRILS
NABA ECM REGULATORS
POSITIVE REGULATION OF ENDOTHELIAL CELL MIGRATION
RECEPTOR LIGAND ACTIVITY
STRIATED MUSCLE TISSUE DEVELOPMENT

Check for unrelated factors

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



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

Avoiding Biases

- Create a custom background if applicable
  - Should contain all genes which *could* have been in your hit list
  - May be a compromise, but it's better than nothing
  - Will limit which tools you can run

- Filter your tested gene sets
  - Remove large over powered sets, or sets which are too small to achieve significance (~50 to ~500 is generally about right)
  - Will clean results and improve the stats for the good hits
  - Check the hit gene sets for matches to known problematic sets