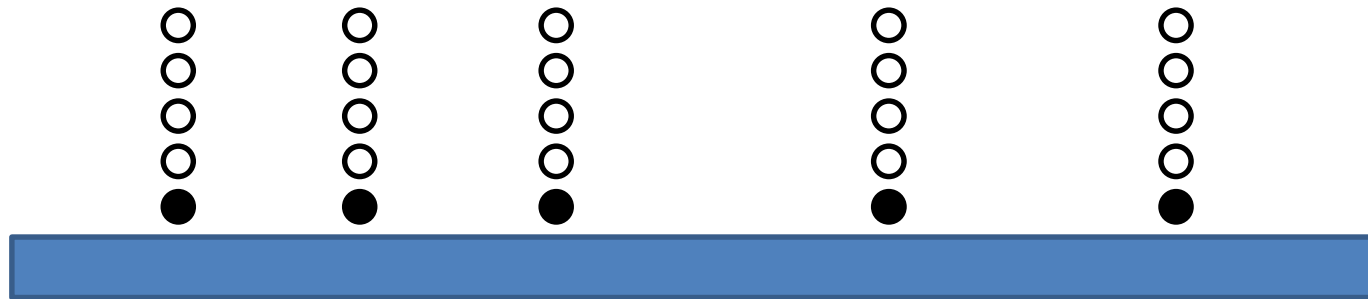
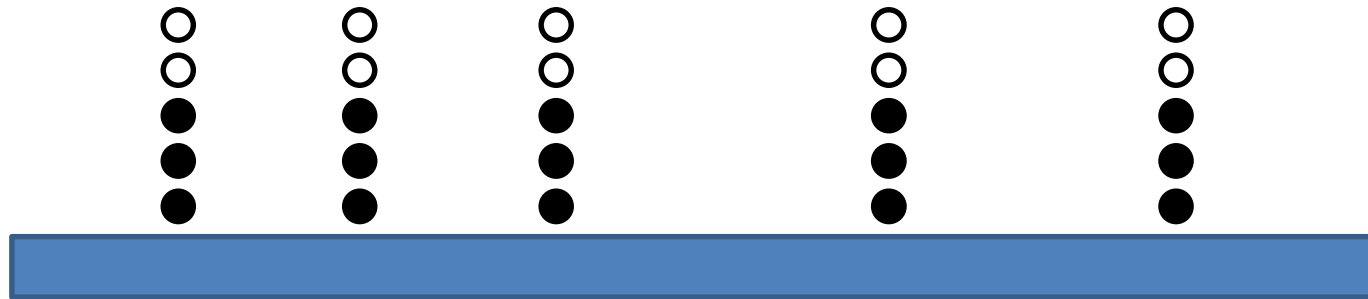


Differential Methylation Analysis

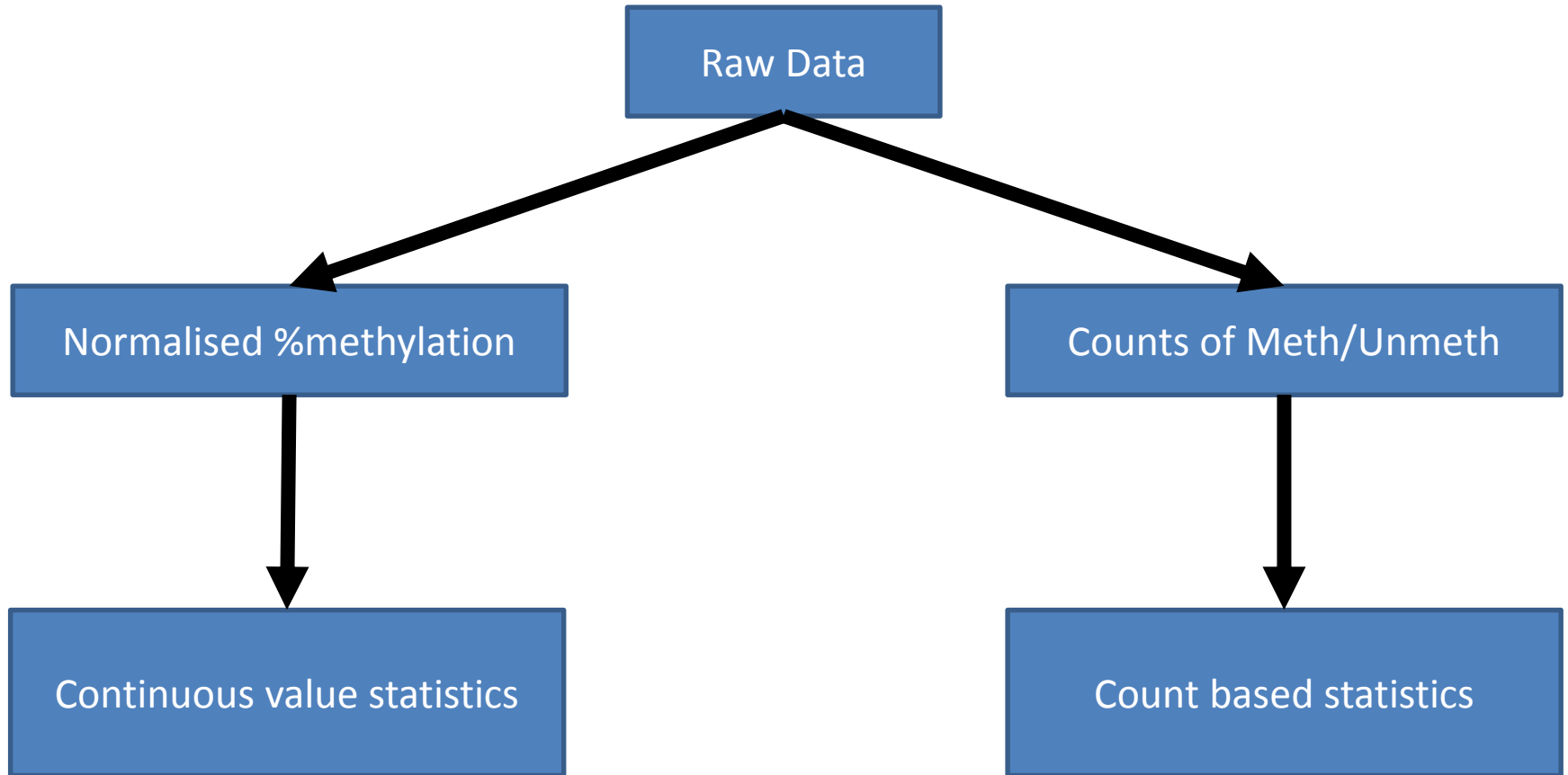
Simon Andrews
simon.andrews@babraham.ac.uk
@simon_andrews

v2017-01

A basic question...



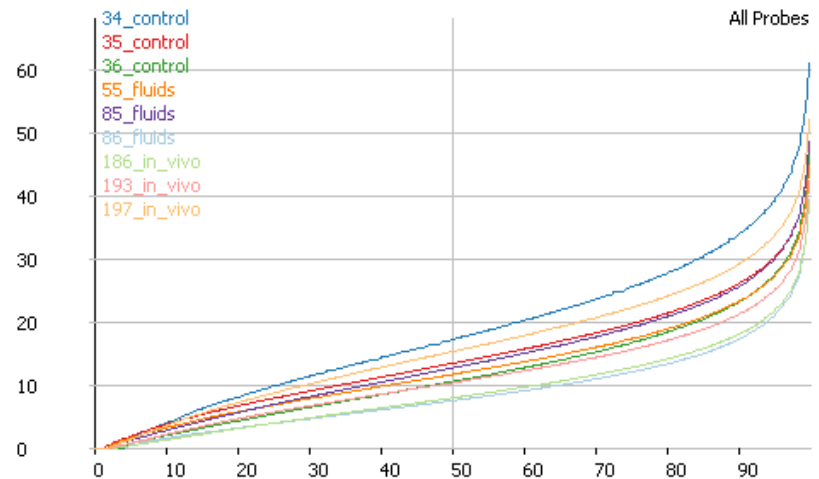
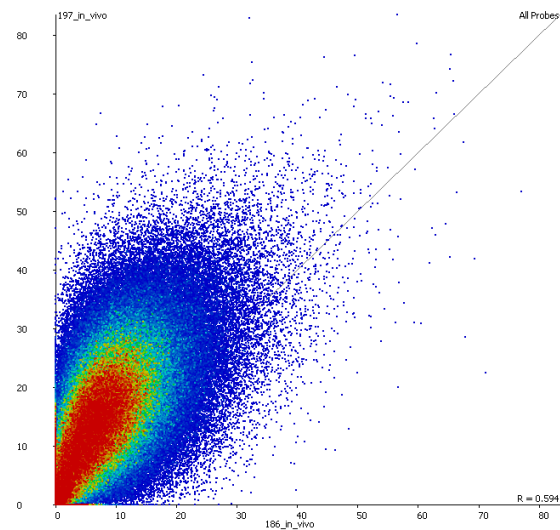
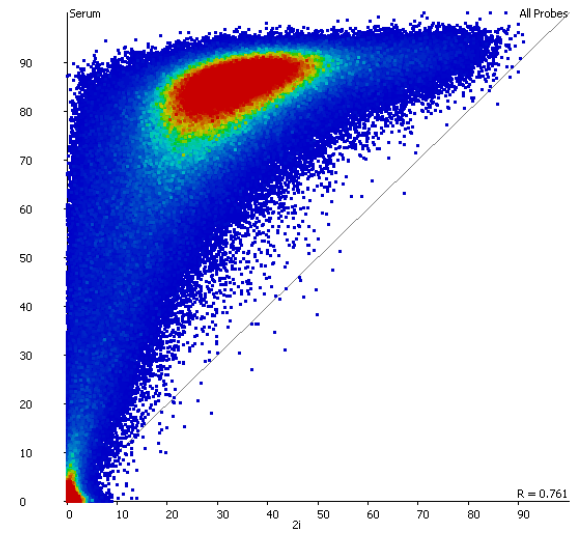
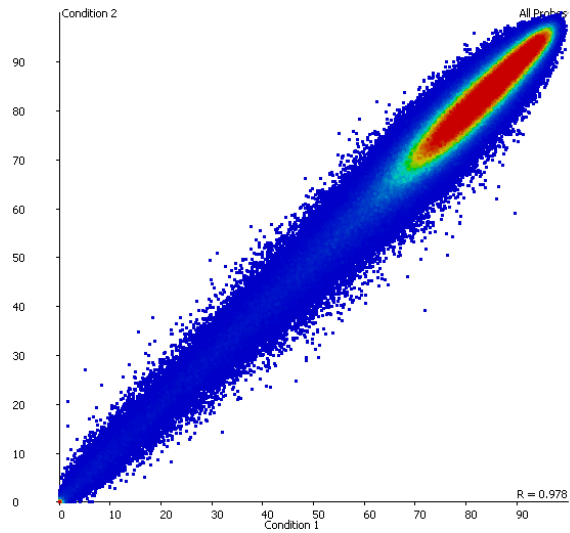
Two Strategies



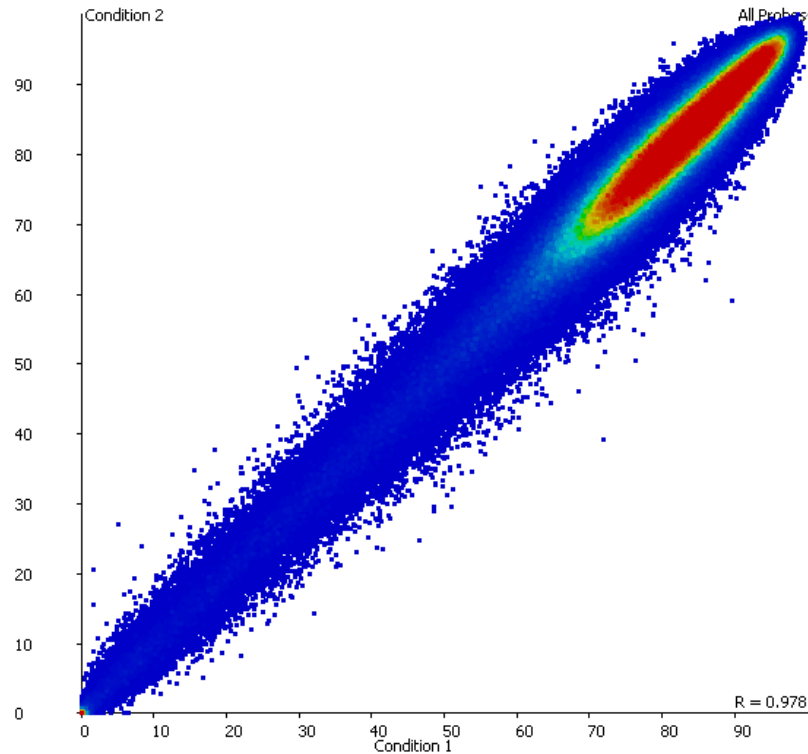
Factors to consider

- Formulating a sensible question
- Applying corrections if needed
- Assessing statistical power
- Relating hits to biology

Global Differences

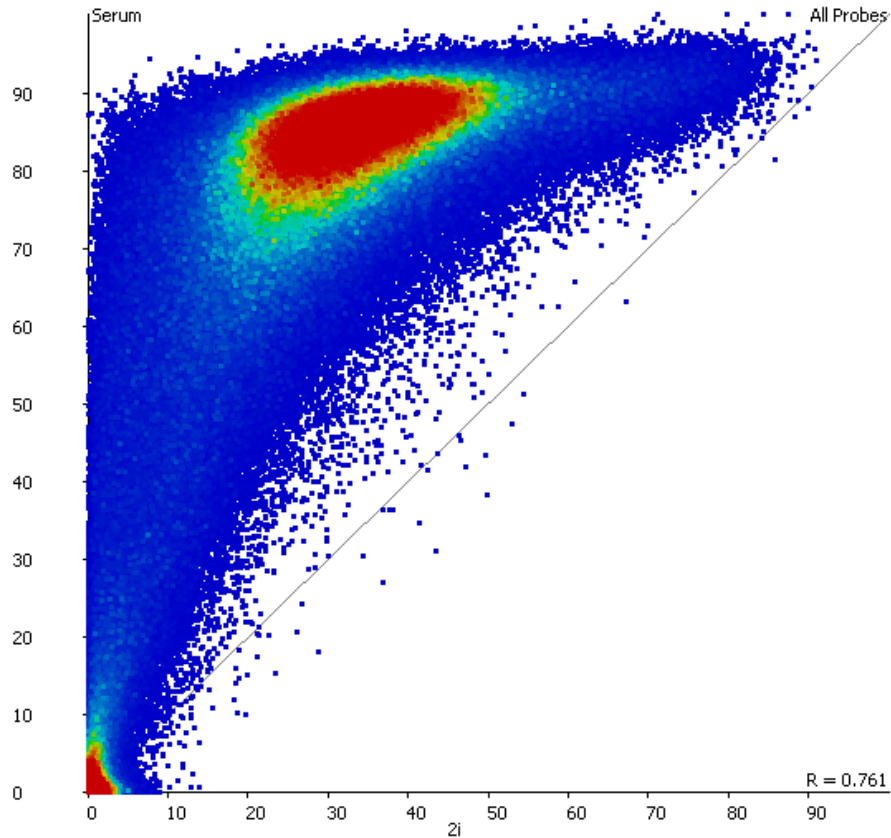


Question



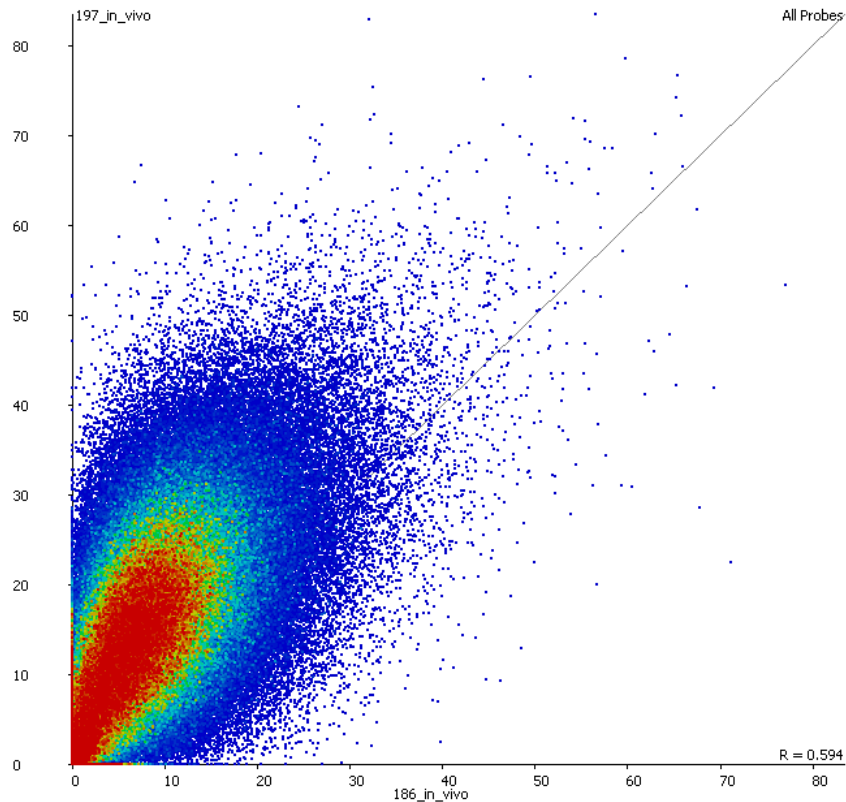
Which areas show a significant change in methylation level between the two conditions?

Question



Which areas show a change in methylation which is larger or smaller than the global change in the samples overall?

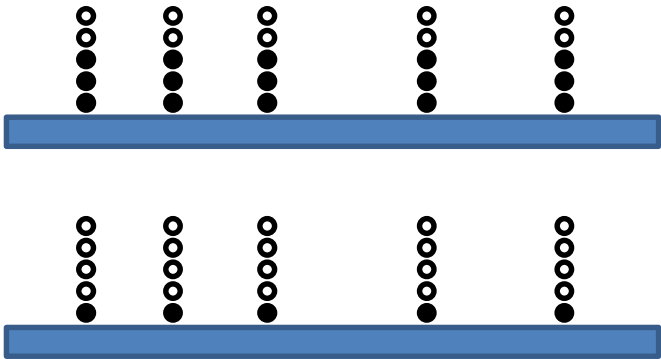
Question



Which areas show a change in methylation after correcting for the small global differences?

Count based statistics

Count Data



	Meth	Unmeth
Sample 1	18	10
Sample 2	5	20

Is the difference in ratios significant given the observation levels of the samples

The problem of power...

- Ideally want to cover every Cytosine (CpG)
- Should correct for the number of tests
- It's unlikely you'll collect enough data to analyse each C and have p-values which survive multiple testing correction
- How do you work around this?

Maximising power

- Options
 - Don't do any multiple testing correction
 - Increase Power
 - Analyse in windows (more data and fewer tests)
 - Pre-filter (fewer tests)
 - Hierarchical or Adaptive filtering

Window sizes

Effect size

Small



Large

- Good resolution
- Specific biological effects
- High MTC burden
- Small observations
- High p-values

- Lots of data
- High statistical power
- Low MTC burden
- Low p-values
- Effect averaging

Power Analysis

(Assuming a human genome with $p < 0.05$ and power of detection of 0.8)

Window Size (# CpG cytosines)

	1	10	25	50	100	200	500
1	158805	14212	5419	2609	1254	602	228
5	6794	608	232	112	54	26	10
10	1825	164	63	30	15	7	3
20	509	46	18	9	5	2	1
50	94	9	4	2	1	1	1

Absolute
methylation
change
(from 80%)

Required Fold Genome Coverage

Without Multiple
Testing Correction

	1	10	25	50	100	200	500
1	25583	2559	1024	512	256	128	52
5	1094	110	44	22	11	6	3
10	294	30	12	6	3	2	1
20	82	9	4	2	1	1	1
50	15	2	1	1	1	1	1



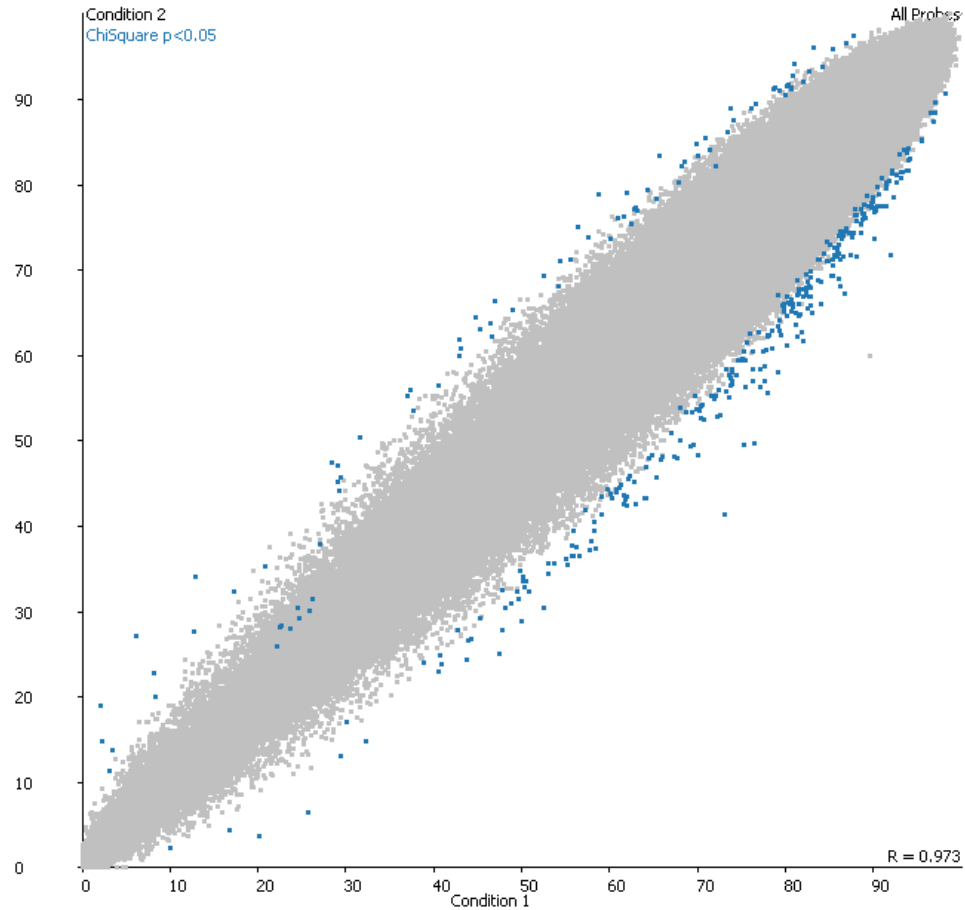
Pre-filtering

- Uncontentious
 - Low coverage regions
 - Regions of biological interest (externally decided)
- Contentious
 - Minimum absolute difference

Contingency tests

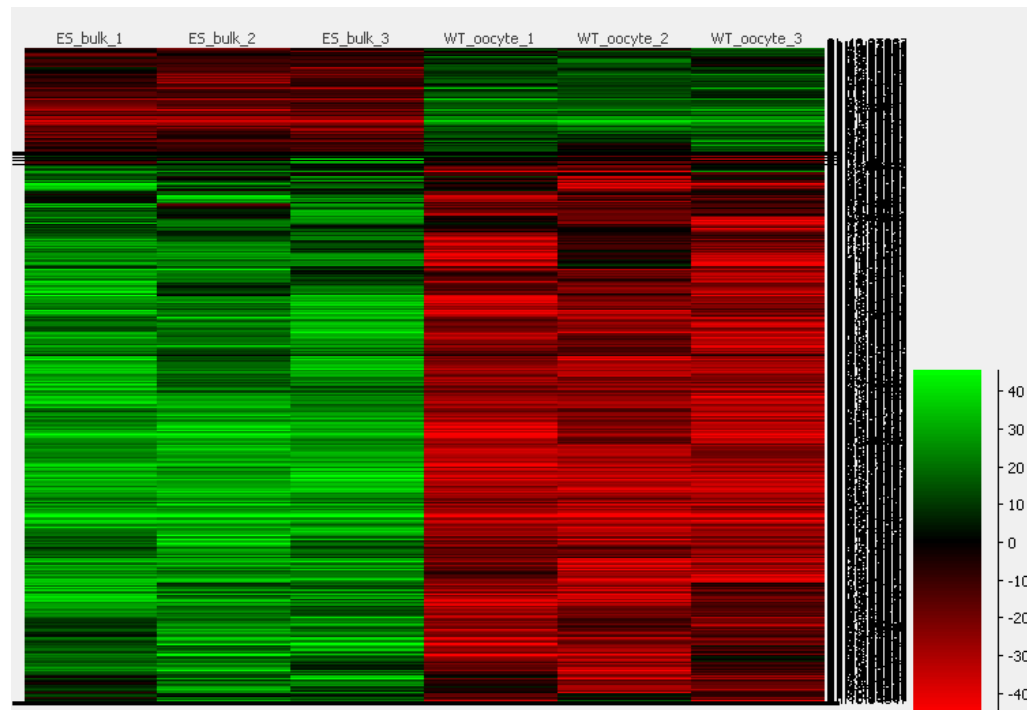
- Chi-square / G-test / Fisher's exact test
 - Differ only at low observations
 - Significant changes require enough observations that any of these will likely give a similar answer
- Operates on single replicates
- Technical measure of difference
- Independent for each measurement window

Chi-Square results

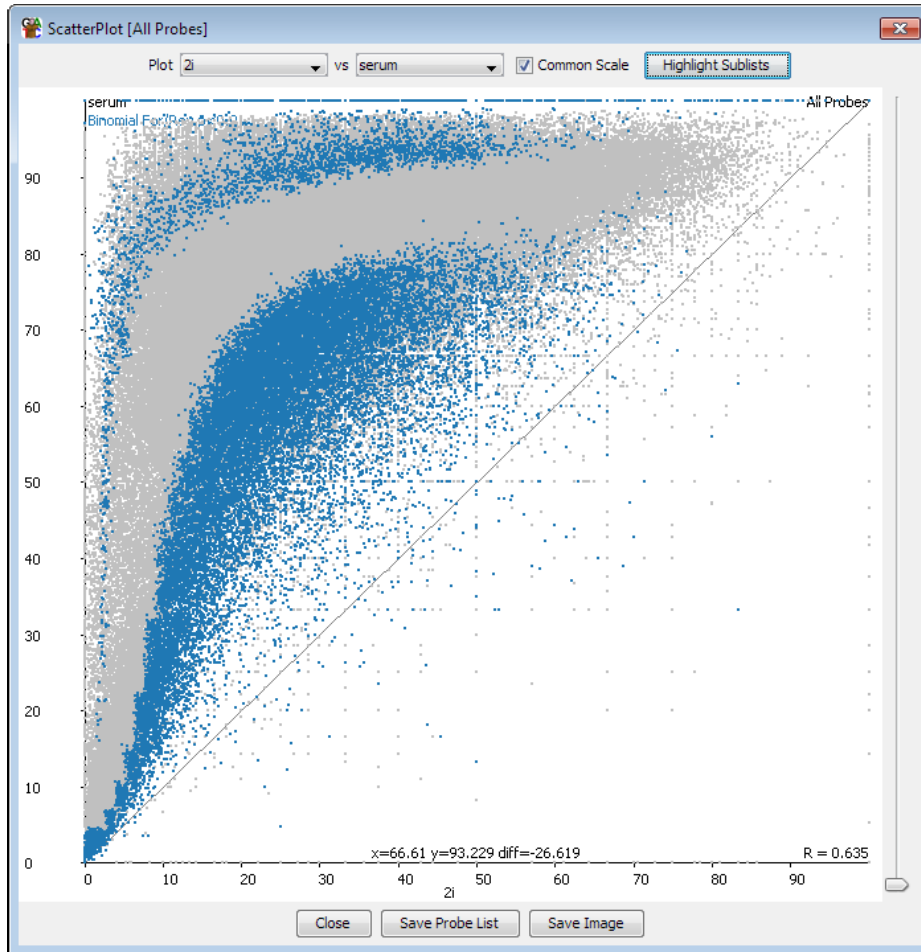


Logistic Regression

- Extension of contingency tables to replicated data
 - Accounts for variability within each sample group
 - Selects points with consistently different methylation given their coverage

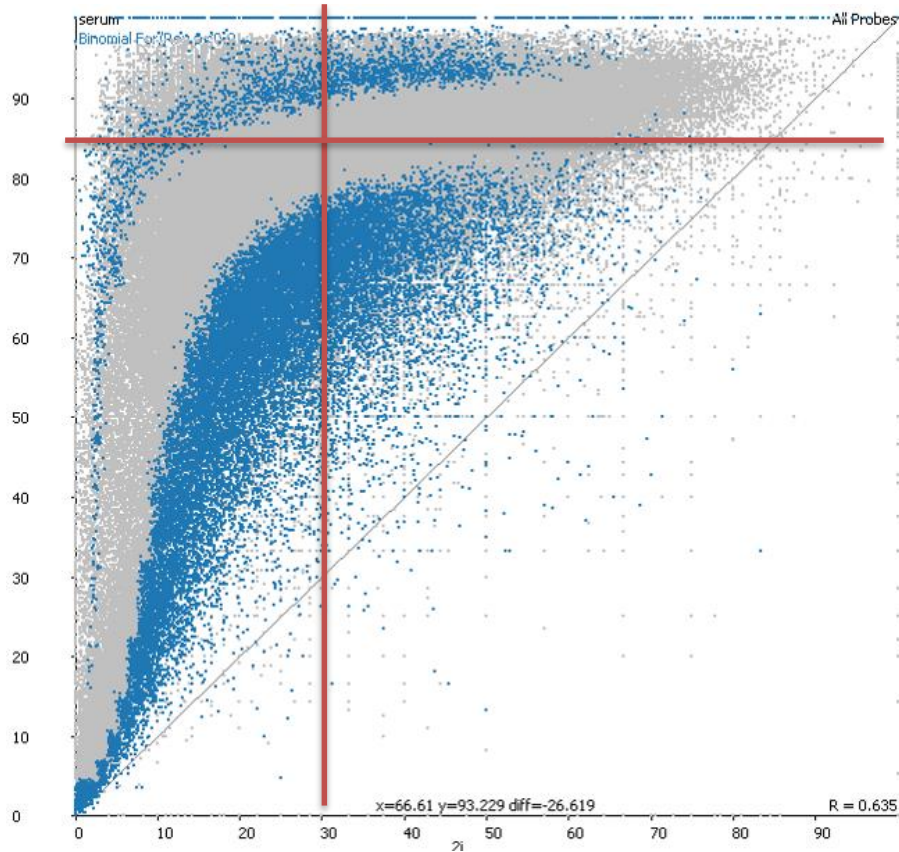


Globally changing samples



- Changes the default expectation
- Find average difference for each starting point
- Select points which exhibit unusual change

Globally changing example



Starting level = 30%

Observations = 14 meth 6 unmeth

Expected End level = 85%

Binomial test, $p=0.85$, trials=20, successes=14

Raw $p=0.106$

Beta Binomial Models



What is the probability distribution for the true methylation level?

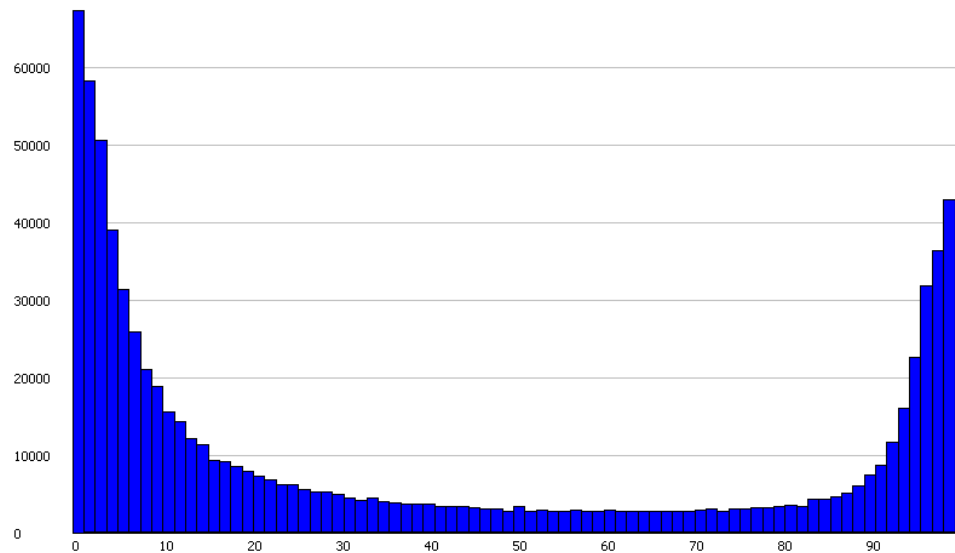
Simple model: Binomial stats to estimate confidence

Can we do better?

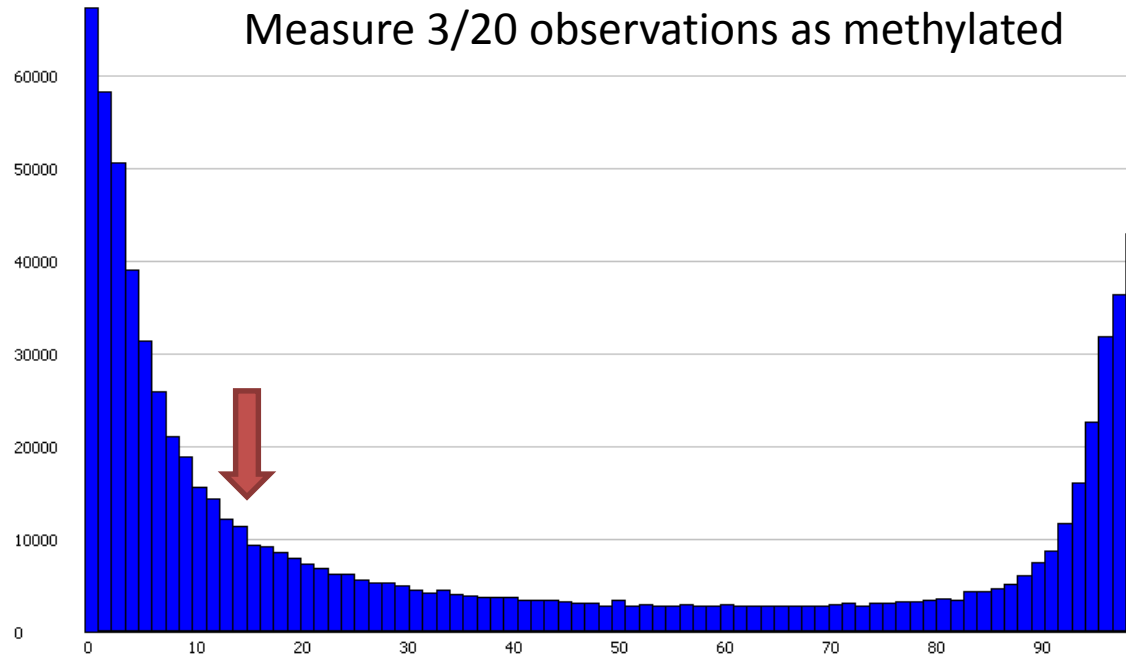
Genome-wide methylation profile.

All levels are not equally likely

Can inform the construction of a
Custom beta binomial distribution



Beta-binomial model



The binomial distribution would be defined by the mean and observations

Using the whole genome prior a beta-binomial model would upweight the lower methylation levels, since these are more common.

Provides increased power in comparisons between major groups

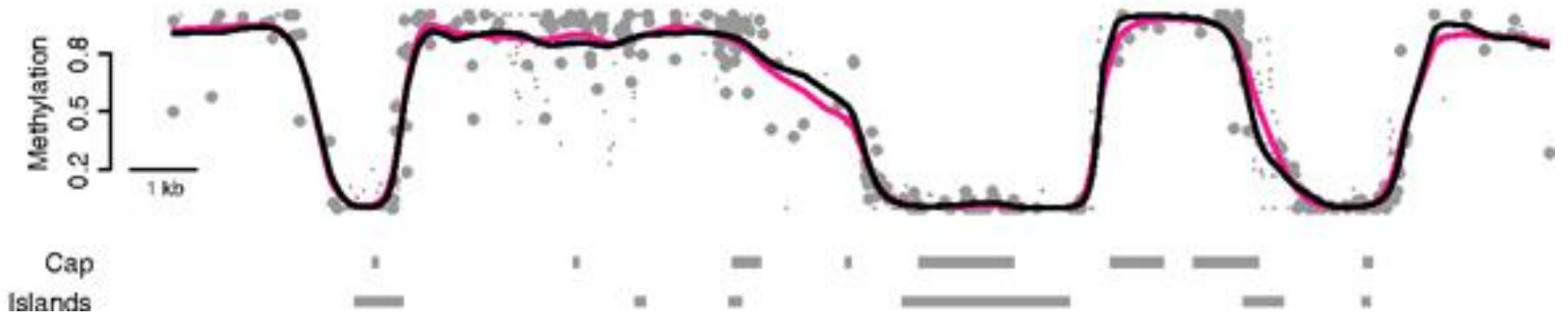
Often computationally intensive

Limitations of count based stats

- No subdivision of calls – all calls are equal even when coverage isn't
 - Supplement with differences based on better quantitation
- Potential biased by power
 - Can alleviate with CpG window based analysis
 - Easy to bias data otherwise
 - Problem of interpretation, not statistics

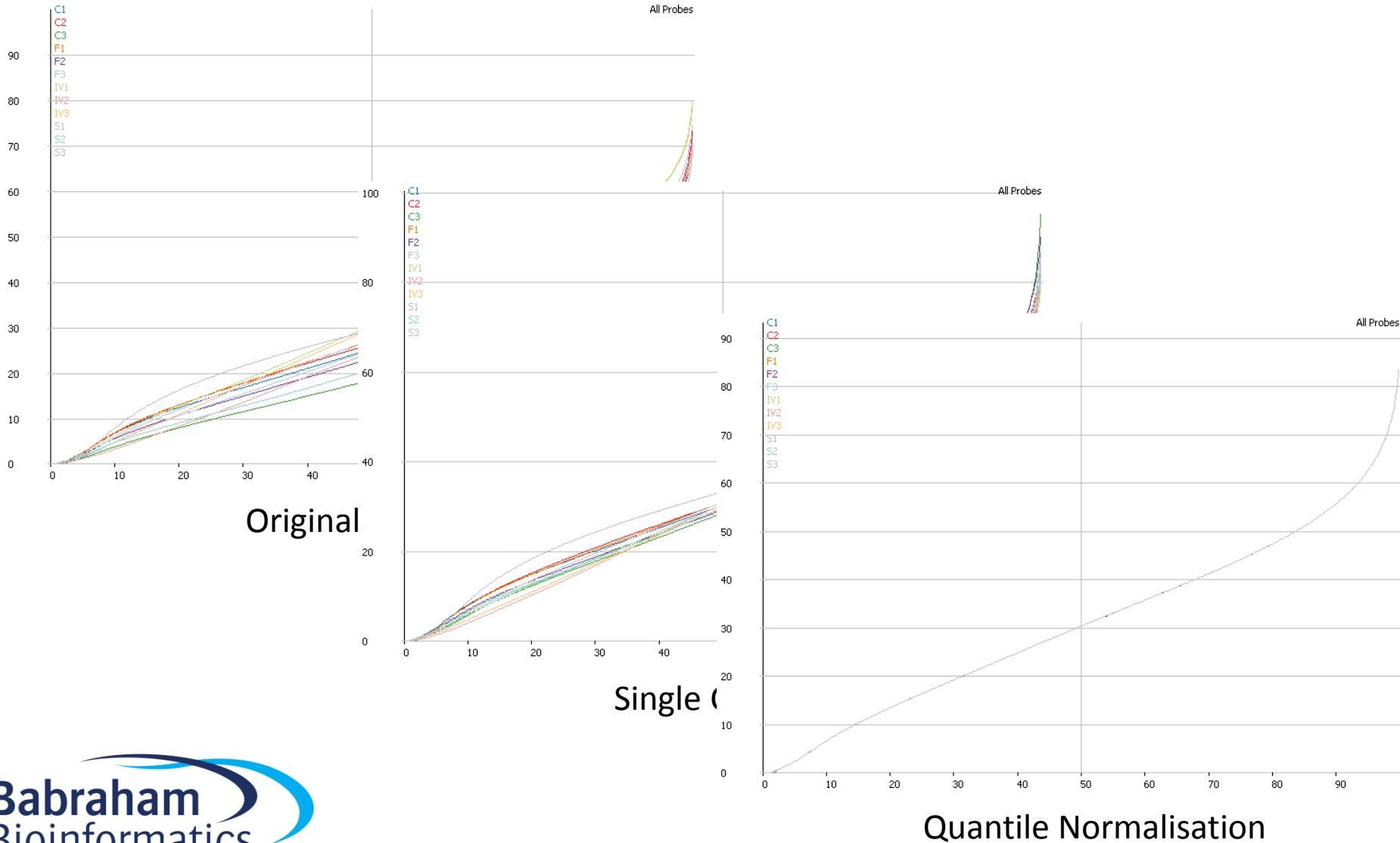
Methylation Level Statistics

BSmooth algorithm for methylation correction



black: 25x (Lister)
pink: 4x (Lister)

Normalisation for methylation levels



Statistics

- Standard continuous statistics
 - T-Test
 - ANOVA
- Much reduced power – one value per replicate
- Unlikely to justify multiple testing correction

Reverse counting

- Some packages offer a conversion from normalised methylation back to counts

True observations: Meth=20 Unmeth=30 (40% meth)

Corrected % methylation = 50%

Reversed counts: Meth=25 Unmeth=25

- Allows count based statistics – regains the lost power from normalisation
- Retains information about noise from true observation level

Biological considerations

- Minimum relevant effect size?
 - Balance power vs change
 - What makes biological sense
 - (what would you follow up?)
- Position relative to features
- Consistent change over adjacent regions

Methylation statistics packages

- **SeqMonk** (Graphical Analysis Package)
Flexible measurement based on fixed windows, fixed calls or features. Complex corrected methylation calculation and several optional post-calculation normalization options. Chi-Square with optional resampling for unreplicated data, logistic regression with optional resampling for replicated data.
- **methylKit** (R-package by A. Akalin et al.)
Sliding window, Fisher's exact test or logistic regression. Adjusts p-values to q-values using SLIM method.
- **bsseq** (R/Bioconductor by K.D. Hansen)
Implements the BSmooth smoothing algorithm. Numerous CpG-wise t-tests and p-value cutoff to define DMRs. Outperforms Fisher's exact test. Requires biological replicates for DMR detection
- **BiSeq** (R/Bioconductor by K. Hebestreit et al.)
Beta regression model, impractical for very large data other than RRBS or targeted BS-Seq
- **MOABS** (C++ command line tool by D. Sun et al.)
Beta binomial hierarchical model to capture sampling and biological variation, Credible Methylation Difference (CDIF) single metric that combines biological and statistical significance