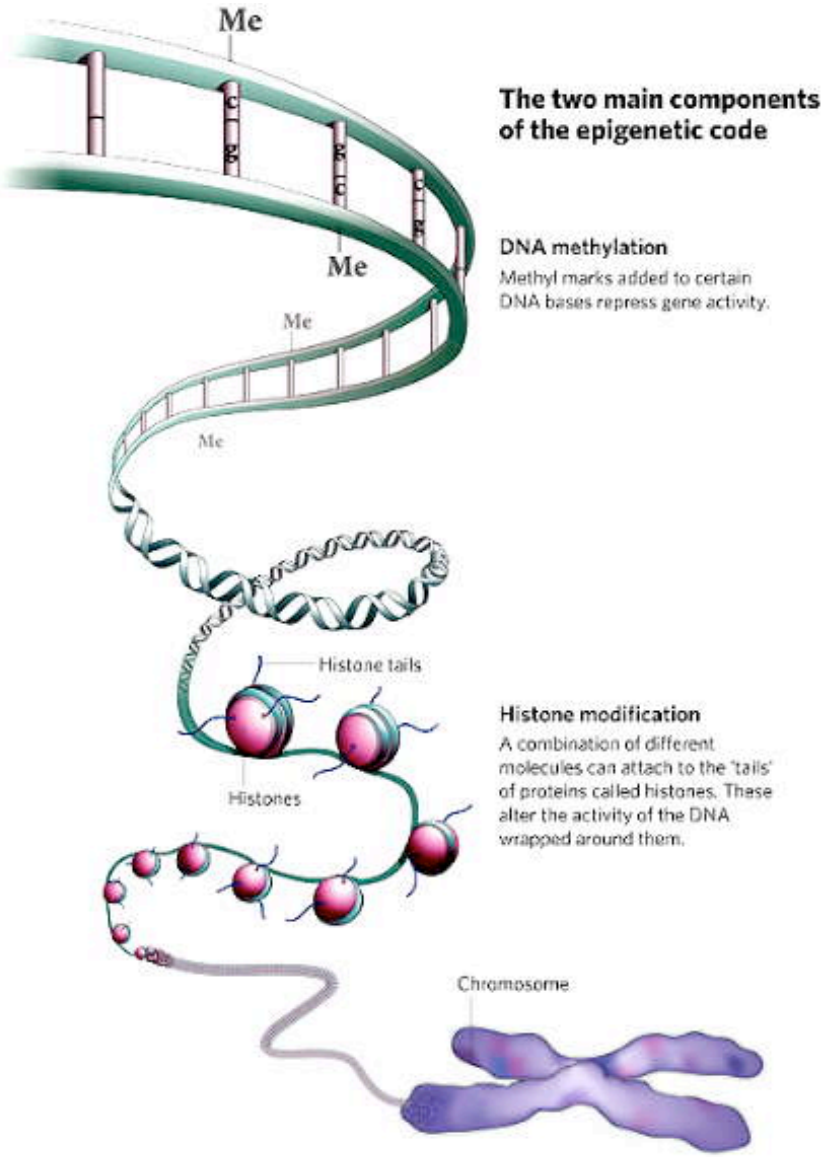
Analysis of DNA methylation microarray data: Epigenome-wide association studies

Karen Conneely
Emory University
Department of Human Genetics

Overview: Epigenetics

- Epigenetics – "above" genetics
  - External changes to DNA that do not alter sequence
  - Epigenetic mechanisms can induce variable expression between genetically identical organisms or cells
  - DNA methylation is one epigenetic mechanism
  - Some others are chromatin remodeling, histone modification (includes methylation, acetylation, phosphorylation, ubiquitination)



Source: Nature Reviews

Epigenetic vs. genetic analysis

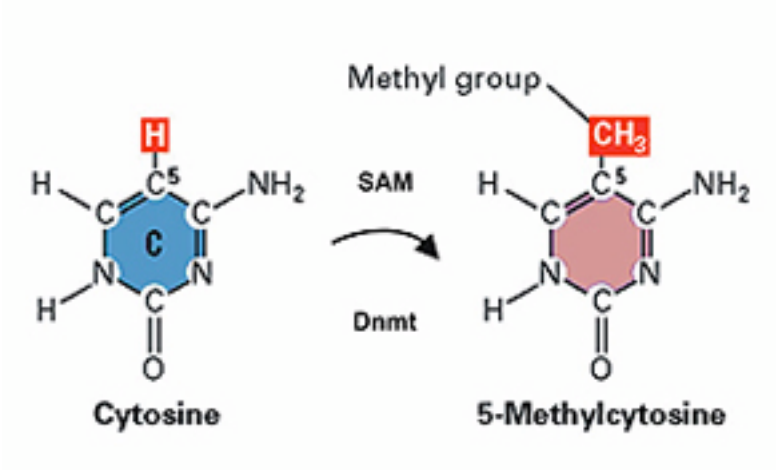
- Genotype is stable (except for somatic mutations, which are rare)
  - Same in every cell of an organism (mitotic inheritance)
  - Stable over time
  - Inherited across generations (meiotic inheritance)
- Epigenotype is dynamic
  - In a single cell, can vary over time
    - Plays role in development (programmed changes)
    - Can be influenced by environment (unprogrammed changes)
  - Varies across cells and tissues
    - Passed on to daughter cells in cell division (mitotic inheritance)
    - Plays important role in cell and tissue differentiation

This lecture: DNA methylation

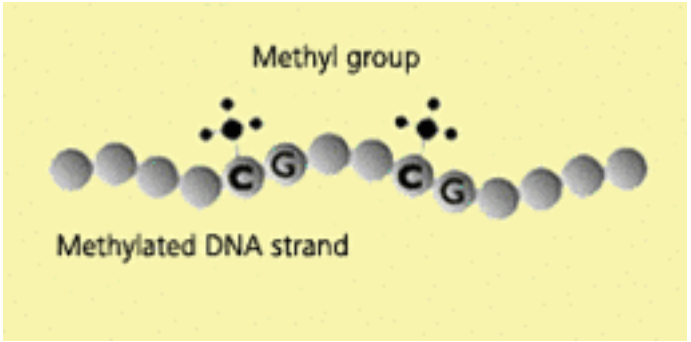
- What is DNA methylation?
- How is it measured?
- How do we analyze methylation data?
- What issues are specific to analysis of these data?
  - Technical factors and data quality
  - Unlike genotype, which is basically static,
    - Methylation differs across tissues and cells
    - Methylation can change over time
  - Interpretation of results
    - Multiple testing adjustment with false discovery rate
    - Test for biological enrichment of results

What is DNA methylation?

- Attachment of a methyl group (CH_3) to a single nucleotide



- Occurs mainly on the C (cytosine) in CpG dinucleotides



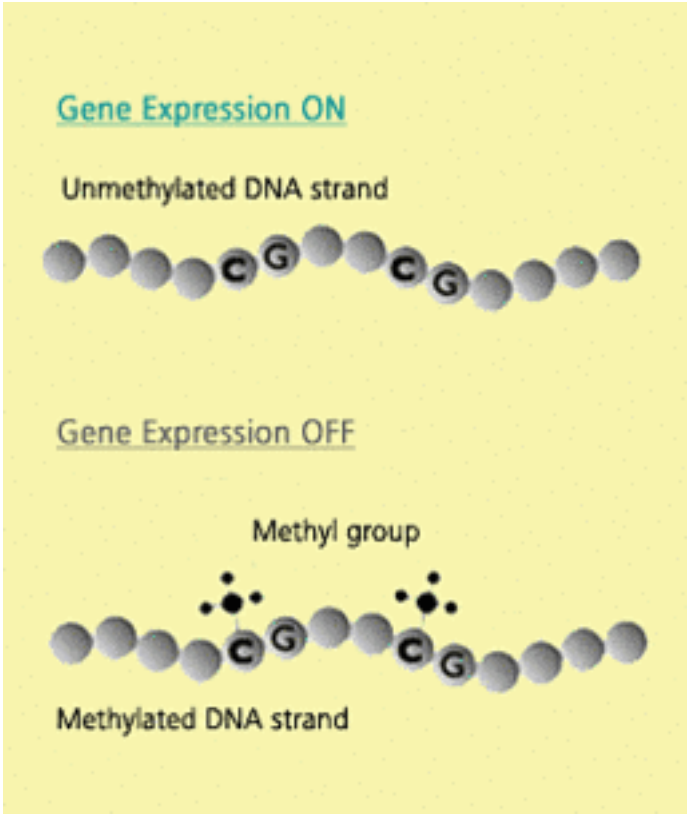
What does it do?

- Can help control gene expression

  Unmethylated DNA is free to be expressed (transcribed as RNA)

  Methylation can silence expression by repressing transcription

- In mammals, most CpG sites across the genome are methylated most of the time
- One exception: CpG-rich regions known as CpG islands
  - Generally located in promoter regions of genes
  - Generally unmethylated
  - Methylation of these regions can silence genes
- Another exception: enhancers
  - Located upstream from promoters
  - Can show variable methylation across cell types (Aran et al. 2013) that controls expression levels

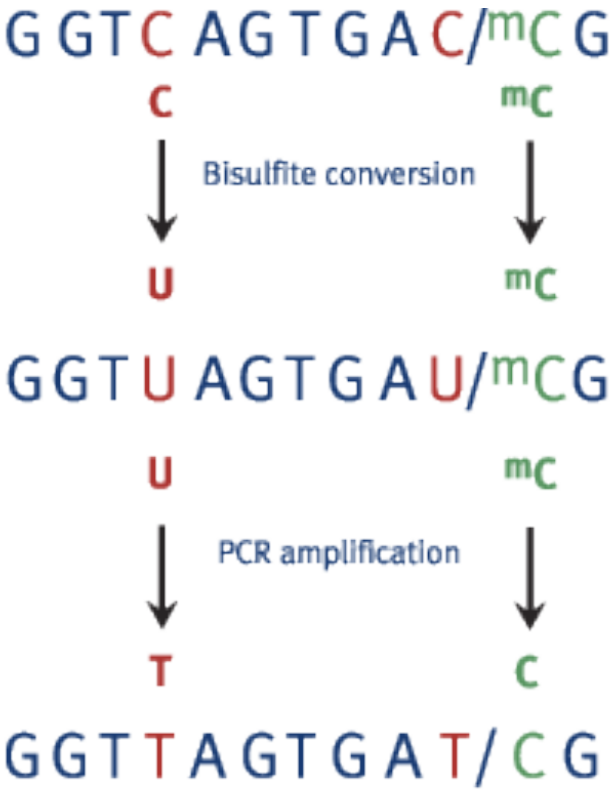


How can DNA methylation be measured?

- Several common approaches
- Site-specific or region-specific methylation
  - Bisulfite treatment of DNA
    - Converts unmethylated Cs to Us
  - Methyl-DNA Immunoprecipitation (MeDIP)
    - Uses an antibody to isolate methylated DNA
  - Restriction enzyme approaches
    - Enzymes that cut only methylated (or unmethylated) DNA
    - Can be used to enrich for methylated (or unmethylated) DNA
  - These approaches can then be combined with microarrays or sequencing to measure methylation at specific locations
- Can also measure global methylation – total methylation across genome

Bisulfite treatment of DNA

- Bisulfite treatment
  - converts unmethylated Cs to Us
  - leaves methylated Cs (mC) as Cs
- PCR amplification
  - Us amplified as Ts
  - Each CpG site now like a C-T SNP
  - Important difference: an individual can have C on some transcripts, T on others
  - Need to assess proportion C (methylated) vs. T (unmethylated)
- Bisulfite-treated DNA can then be sequenced or “genotyped” on arrays



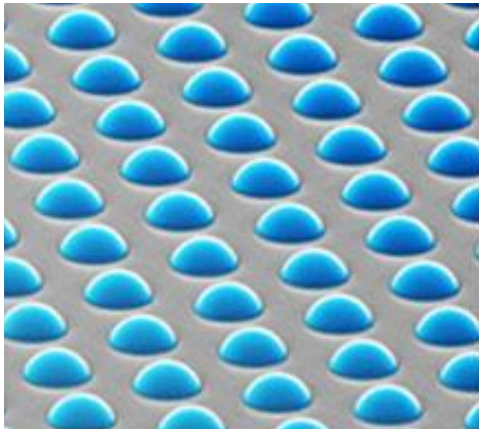
Genome-wide methylation arrays

- Genome-wide arrays allow hypothesis-free (as opposed to candidate gene) approach to methylation studies
- Availability of commercial DNA methylation arrays has made genome-wide methylation studies feasible for many labs
- Right now, "commercial" methylation arrays = Illumina
  - 2007: Goldengate array; ~1500 CpG sites in cancer-relevant genes
  - 2009: Infinium 27K; ~27K CpG sites representing ~14K genes
  - 2011: Infinium 450K; >480K CpG sites covering 99% of RefSeq genes and 96% of CpG islands in UCSC database[1]

[1]Bibikova et al. (2011) *Genomics* 98:288-95

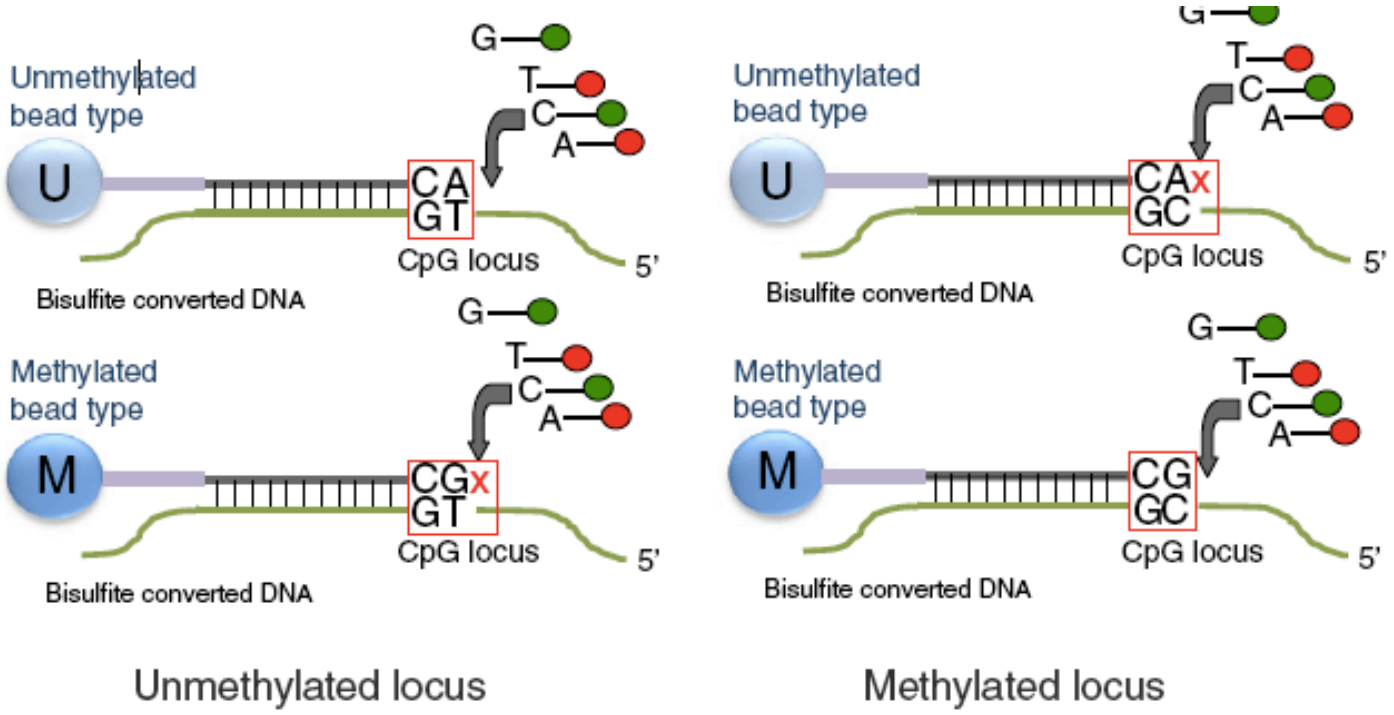
Infinium 450K array

- Bead-bound probes assay methylation at 485,577 CpG sites and other probes
  - DNA is bisulfite treated to convert unmethylated C to T
  - Whole genome amplification is performed
  - DNA fragments anneal to 50-mer probes
  - Single base extension used to measure signals for methylated (M) and unmethylated (U) DNA
- What does the resulting data look like?
  - Methylation "β value" computed as ratio of methylated to total signal



$$\beta = \frac{M}{M + U}$$

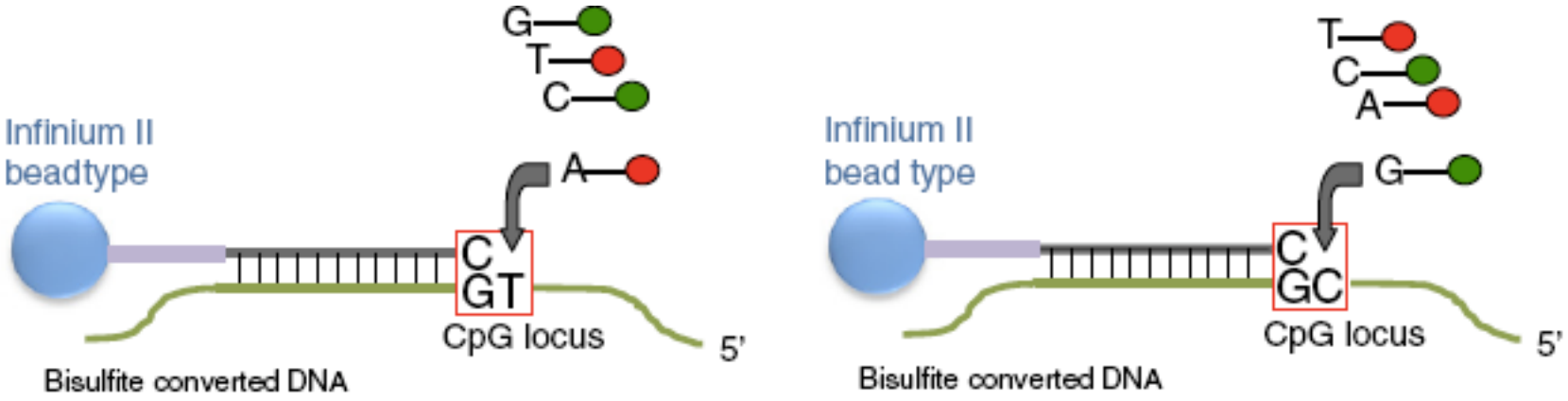
Infinium Type I assay



– Each bead type detects presence of C or T by hybridization followed by single-base extension with a labeled nucleotide
  - Methylated (M) bead type: detects presence of C
  - Unmethylated (U) bead type: detects presence of T

$$\beta = \frac{M}{M + U}$$

Infinium Type II assay



- Advantage: probes can be "degenerate" at up to 3 positions to allow for additional CpG sites within a probe
- Disadvantage: Possible color channel bias?

$$\beta = \frac{M}{M+U} = \frac{Red}{Red+Green}$$

Bibikova et al. (2011) *Genomics* 98:288-95

Infinium I vs. II assays

- Because of differences in type I vs. II assays, their β-value distributions differ[1,2]
- For analyses that combine multiple probes, can rescale type II β-values to match type I distribution[2]
- For single-CpG analyses, this step may not be strictly necessary

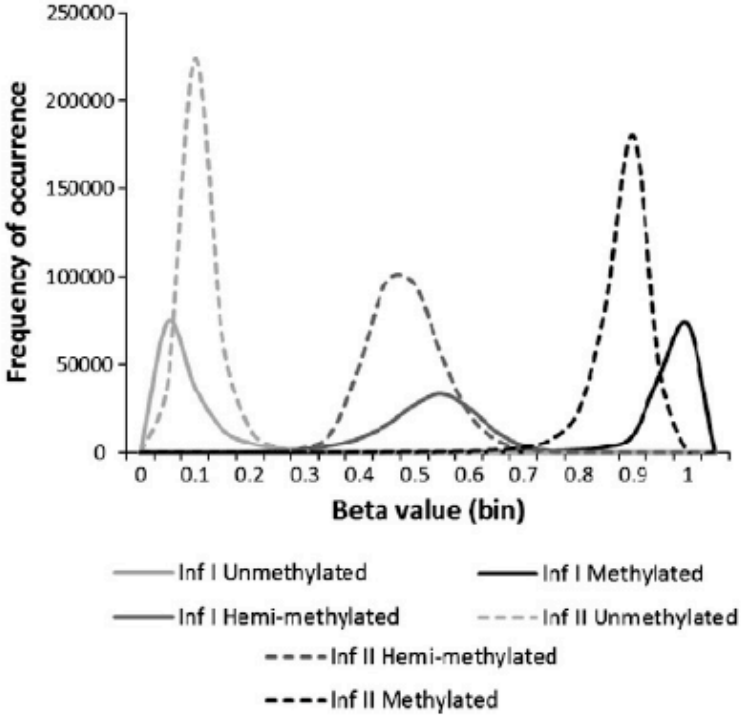


Fig. 3. Distribution of Methylation values for Infinium I and Infinium II loci. Unmethylated (U), Hemi-methylated (H), and Methylated (M) reference standards were created from Coriell genomic DNA sample as discussed in Methods. Note slightly different performance of Infinium I and Infinium II assays in regard to beta value distribution.

[1]Bibikova et al. (2011) *Genomics* 98:288-95
[2]Dedeurwaerder et al. (2011) *Epigenomics* 3:771–84

So what do the Illumina data look like?

- Illumina's GenomeStudio provides 4 *K* × *N* matrices, where *K* = # of CpG sites and *N* = # *of* samples
  - *M* signal
  - *U* signal
  - *β*-value $\beta = \frac{M}{M + U + 100}$
    - Essentially a proportion between 0 and 1
    - The constant 100 is arbitrary and can be dropped
    - e.g., if *M*=5000 and *U*=10,000, β = 0.33
    - For this sample, roughly 33% of DNA strands assayed are methylated
  - Detection p-values: measures of signal quality
    - For each CpG site and sample, tests null hypothesis that observed signal could simply be noise

Steps in analysis of 450K data:

- Pre-processing: quality control and normalization
  - Basic quality control of the data
  - Possible filtering of CpG sites
  - Normalization to remove between-sample differences or within-sample differences between type I and II probes
- Analysis of the data
  - Test for association between methylation at single CpG sites and variable of interest (Epigenome-wide association study = EWAS)
    - Disease or disease-related phenotypes
    - Treatment or intervention
    - Environmental exposure

Quality control of Illumina array data

- Goal: identify/remove poorly performing samples, and use detection p-values to identify noisy data points
- Our R package CpGassoc [1] includes a function to perform basic quality control on Illumina signal data
- For example, the command
  - cpg.qc(beta, signalA, signalB, detection_pval, p.cutoff=.001, cpg.miss=.1, sample.miss=.05)
  - Sets to missing data points with detection P-value>.001
  - Removes CpG sites with >10% missing data
  - Removes samples with >5% missing data or signal < 50% of median

[1]Barfield et al. (2012) *Bioinformatics* 28:1280-1
Available at http://genetics.emory.edu/conneely

Possible ways to filter set of CpG sites

- Removal of non-specific probes
  - List of probes mapping to multiple regions available at[1]
- Remove sites with documented SNP-in-probe
  - Available in Illumina annotation (but incomplete)
  - List based on dbSNP available at[1]
  - List based on 1000 Genomes Project variants with MAF>.01 available at[2]
- Some studies restrict to most variable CpG sites
  - Problem: this can enrich for sites with outliers, or sites under genetic control (mQTLs)

[1]Price et al. (2013) *Epigenetics and Chromatin* 6(1):4
[2]Barfield et al. (2014) *Genetic Epidemiology* 38(3):231-41
Available at http://genetics.emory.edu/conneely or http://genetics.emory.edu/research/?assetID=2161

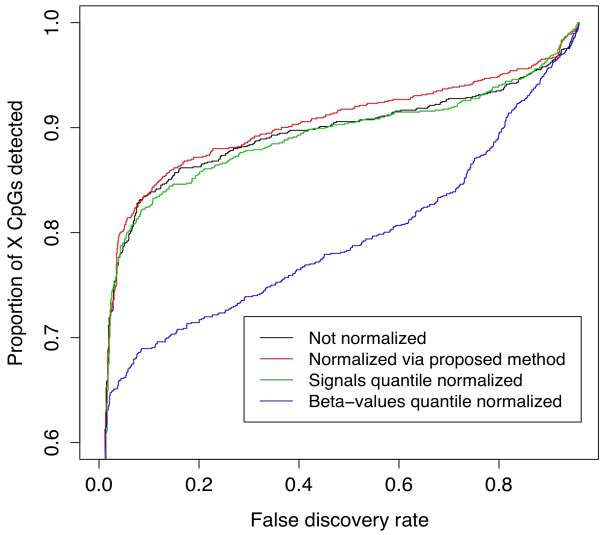
Is normalization necessary?

- For gene expression analyses, yes.
  - These analyses rely on measurements of absolute signal
  - Normalization is thus crucial to remove global differences between samples
- For methylation analyses, less clear.
  - These analyses rely on signal ratios
  - Between-sample differences in signal intensity tend to be cancelled out in the ratio

$$\text{Example: } \frac{2M}{2M+2U} = \frac{M}{M+U}$$

Is normalization necessary?

- "In gene expression analysis, most normalization algorithms operate under the assumption that the majority of genes are not differentially expressed. However, we cannot make this same assumption for the purposes of methylation analysis." (Illumina documentation)
- Nevertheless, quantile normalization is sometimes applied to remove between-sample differences in genome-wide methylation distribution
- If used, should be applied to combined (U and M) signal data, rather than to beta values
- This will minimize removal of global methylation differences
- Extreme example: ability to detect association of X chromosome CpG sites with gender



For example, if A is our signal matrix:

$$\begin{pmatrix} 124 & 588 & 544 & 412 \\ 515 & 712 & 398 & 651 \\ 671 & 423 & 645 & 516 \\ 782 & 814 & 743 & 687 \end{pmatrix},$$

we then order the values for each individual:

$$\begin{pmatrix} 124 & 423 & 398 & 412 \\ 515 & 588 & 544 & 516 \\ 671 & 712 & 645 & 651 \\ 782 & 814 & 743 & 687 \end{pmatrix},$$

take the average across the ordered values:

$$\begin{pmatrix} 339.25 & 339.25 & 339.25 & 339.25. \\ 540.75 & 540.75 & 540.75 & 540.75 \\ 669.75 & 669.75 & 669.75 & 669.75 \\ 756.5 & 756.5 & 756.5 & 756.5 \end{pmatrix},$$

and put the values back in their original order. Thus the quantile normalized data would be:

$$\begin{pmatrix} 339.25 & 540.75 & 540.75 & 339.25 \\ 540.75 & 669.75 & 339.25 & 669.75 \\ 669.75 & 339.25 & 669.75 & 540.75 \\ 756.5 & 756.5 & 756.5 & 756.5 \end{pmatrix}.$$

source: Barfield et al., *Gen Epid* 2014

Some other common normalization strategies

- "Illumina" normalization (an option in GenomeStudio)
  - Measures intensity in ~90 negative control probe pairs with no underlying CpG sites in the probe
  - Signals for all CpG sites in a sample are divided by the average intensity of negative controls in that sample
- BMIQ – Beta MIxture Quantile dilation[1]
  - Uses 3-state mixture model to assign each probe to one of three categories (methylated, hemi-methylated, unmethylated)
  - Within each category, normalizes type II probes to match distribution of type I probes
- SWAN – Subset Quantile Within-Array Normalization[2]
  - Quantile normalization based on a subset of type I and II probes determined to be "biologically similar" based on CpG content
  - Brings type I and II distributions closer together

[1]Teschendorff et al., *Bioinformatics* 2013
[2]Maksimovic et al., *Genome Biology* 2012

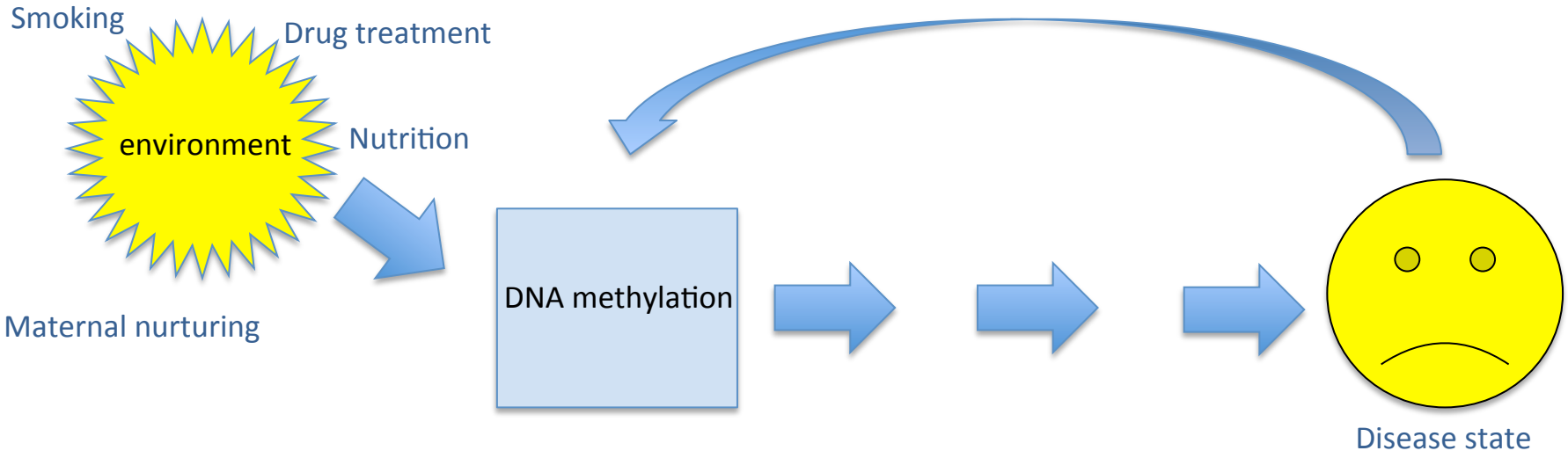
Comparison of normalization strategies

- A recent study compared the reproducibility of different normalization strategies in 52 technical replicate samples[1]
  - 13 pairs (technical duplicates)
  - 2 samples replicated 13 times each
- "Remarkably, the raw, un-normalized data are already highly reproducible and the improvements offered by BMIQ and SWAN are modest "
- Other approaches (including Illumina) can actually introduce additional variability into the data
- When an EWAS of cotinine levels is repeated using each method, results are very similar for all normalization approaches

[1]Wu et al., *Epigenetics* 2014

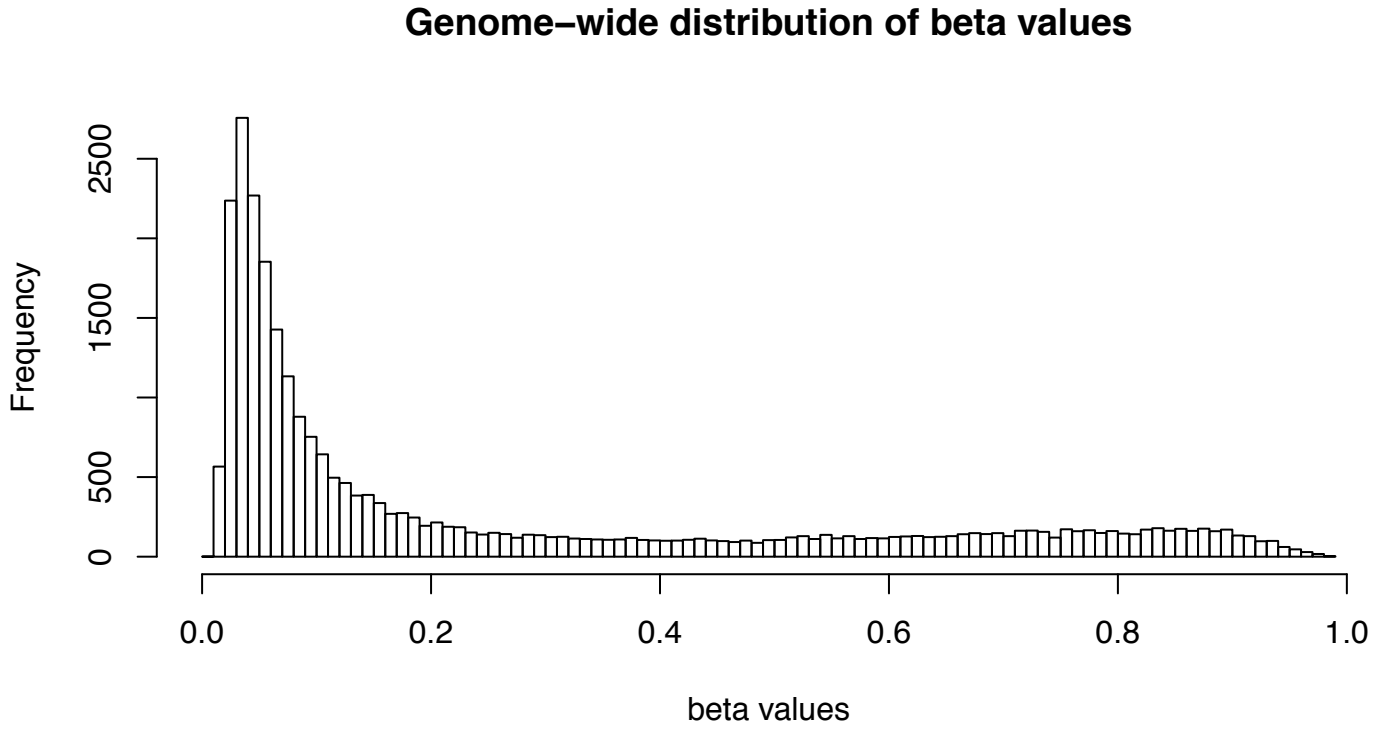
Issues in analysis of methylation array data

- Data quality control and normalization
- Analysis of the data
  - Test for ***association*** between methylation at single CpG sites and variable of interest
    - Disease or disease-related phenotypes
    - Treatment or intervention
    - Environmental exposure



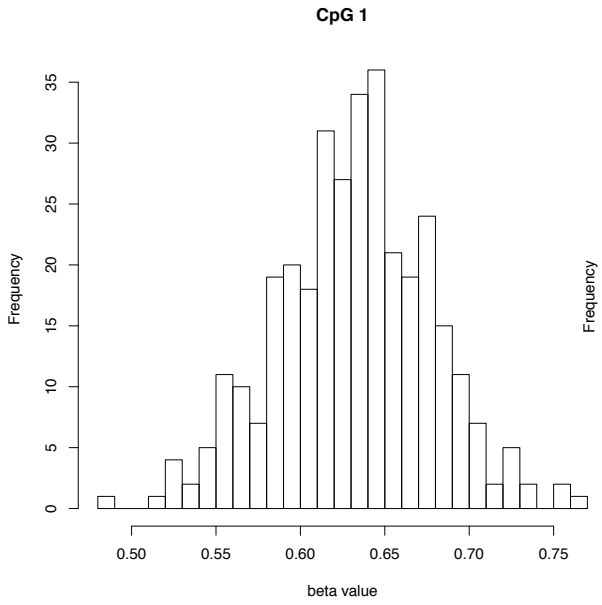
Testing for association

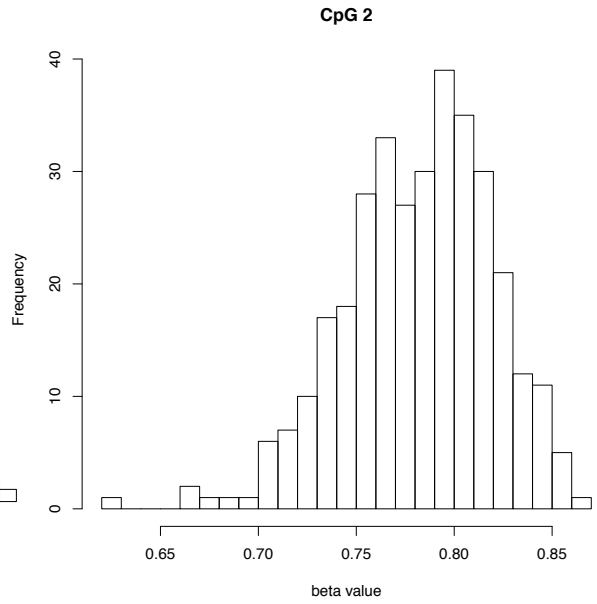
- In an EWAS, test each CpG site for association separately
- Common misconception: methylation has a bimodal distribution, so standard regression techniques are invalid
- It is true that methylation is distributed bimodally ***across the genome***

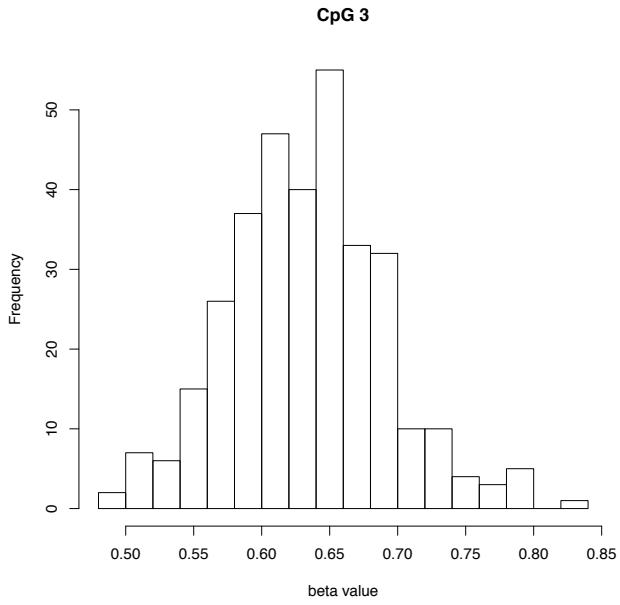


Testing for association

- In an EWAS, test each CpG site for association separately
- Common misconception: methylation has a bimodal distribution, so standard regression techniques are invalid
- It is true that methylation is distributed bimodally ***across the genome***
- However, at each CpG site, distribution of beta values ***across individuals*** is typically unimodal and close to normal*







Testing for association

- In an EWAS, test each CpG site for association separately
- Common misconception: methylation has a bimodal distribution, so standard regression techniques are invalid
- It is true that methylation is distributed bimodally ***across the genome***
- However, at each CpG site, distribution of beta values ***across individuals*** is typically unimodal and close to normal*
  - * After filtering out CpG sites harboring SNPs
- Thus, it is appropriate to use regression-based methods for CpG-level analyses

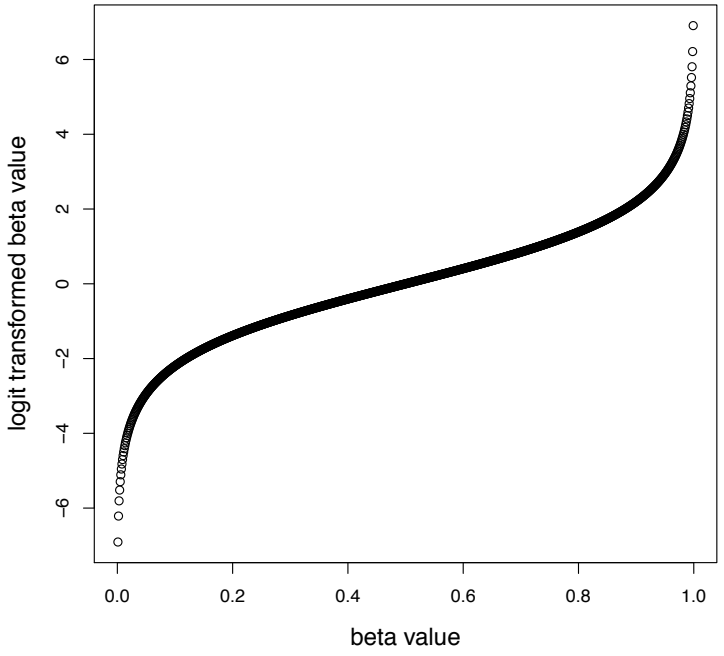
Testing for association

- In an EWAS, test each CpG site for association separately
- Common to use some kind of regression analysis, where methylation *β*-value is the outcome

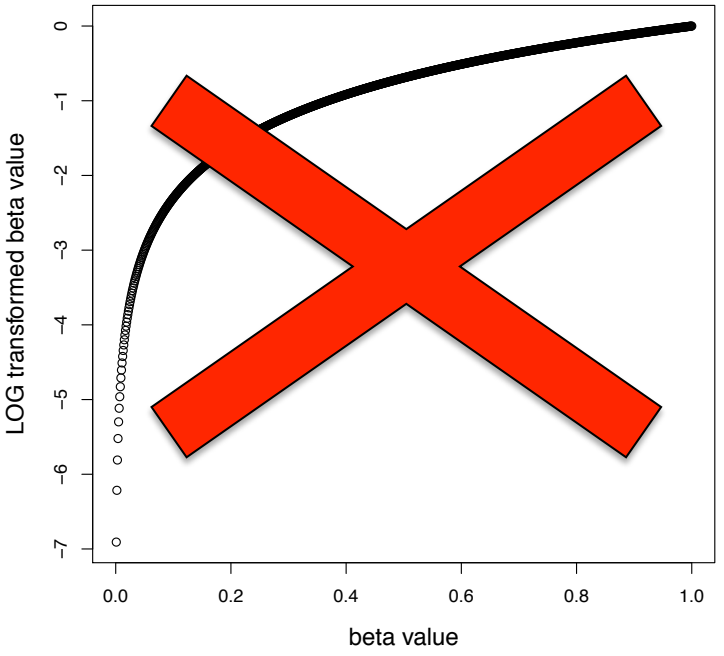
Logit transform or M-value[1]

$$\beta_i = \alpha_0 + \alpha_1 \cdot X_i + \varepsilon_i \qquad \text{or} \qquad \log\left(\frac{\beta_i}{1-\beta_i}\right) = \alpha_0 + \alpha_1 \cdot X_i + \varepsilon_i$$

Logit transform



Log transform



[1]Du et al. BMC Bioinformatics 2010

Testing for association

- In an EWAS, test each CpG site for association separately
- Common to use some kind of regression analysis, where methylation *β*-value is the outcome

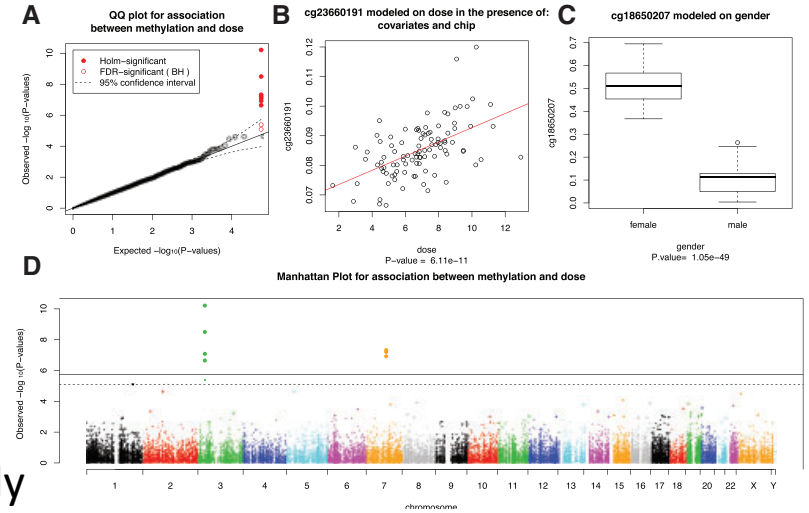
$$\beta_i = \alpha_0 + \alpha_1 \cdot X_i + \varepsilon_i \qquad \text{or} \qquad \log\left(\frac{\beta_i}{1-\beta_i}\right) = \alpha_0 + \alpha_1 \cdot X_i + \varepsilon_i$$

- Most analyses will require additional covariates
  - 1. Technical variation (Batch/chip/positional effects)
  - 2. Sex
  - 3. Ancestry/race (population stratification)
  - 4. Cell type composition
  - 5. Age

$$\beta_i = \alpha_0 + \alpha_1 \cdot X_i + \alpha_2 \cdot \mathrm{cov}_1 + \alpha_3 \cdot \mathrm{cov}_2 + \varepsilon_i$$

Analysis software for EWAS

- To address the need for user-friendly and efficient software to analyze methylation data, we made available two R packages
  - CpGassoc [1] – a suite of R functions
  - MethLAB [2] – a GUI with a menu-driven format
- These packages allow users with little or no R experience to
  - fit fixed and mixed effects models to model methylation as a function of phenotype of interest and other relevant covariates
  - analyze large datasets rapidly
  - summarize results with plots



[1] Barfield et al. (2012) *Bioinformatics* 28:1280-1
[2] Kilaru et al. (2012) *Epigenetics* 7:225-9
Available at http://genetics.emory.edu/conneely

Analysis software for EWAS

CpGassoc provides a standard R interface for a suite of functions

cpg.assoc package:CpGassoc R Documentation

Association Analysis Between Methylation Beta Values and Phenotype of Interest

Usage:

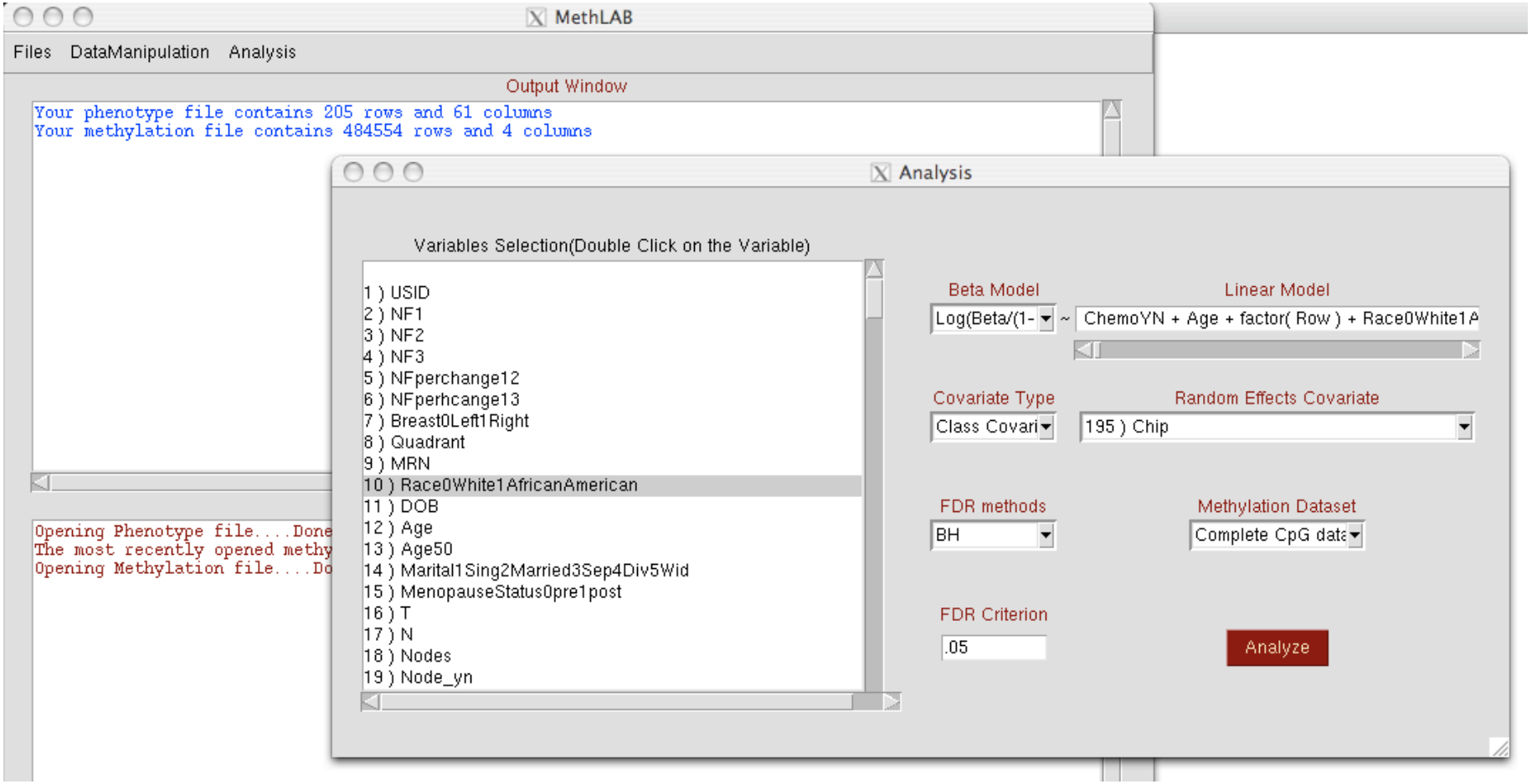
```
cpg.assoc(beta.val, indep, covariates = NULL, data = NULL, logit.transform = FALSE, chip.id = NULL, subset = NULL, random = FALSE, fdr.cutoff = 0.05, impute = FALSE, large.data = FALSE, fdr.method = "BH", logitperm = FALSE)
```

Arguments:

```
beta.val: A vector, matrix, or data frame containing the beta values of
          interest (1 row per CpG site, 1 column per individual).

   indep: A vector containing the variable to be tested for
          association. cpg.assoc will evaluate the association between
          the beta values (dependent variable) and indep (independent
          variable).

covariates: A data frame consisting of additional covariates to be
          included in the model. covariates can also be specified as a
          matrix if it takes the form of a model matrix with no
          intercept column, or can be specified as a vector if there is
          only one covariate of interest. Can also be a formula(e.g.
          ~cov1+cov2).

    data: an optional data frame, list or environment (or object
          coercible by as.data.frame to a data frame) containing the
          variables in the model. If not found in data, the variables
          are taken from the environment from which cpg.assoc is
          called.

logit.transform: logical. If TRUE, the logit transform of the beta
          values log(beta.val/(1-beta.val) will be used. Any values
          equal to zero or one will be set to the next smallest or
          largest value respectively.

 chip.id: An optional vector containing chip or batch identifiers. If
          specified, chip id will be included as a factor in the model.

  subset: An optional logical vector specifying a subset of
          observations to be used in the fitting process.

:
```

Analysis software for EWAS

...while MethLAB provides an easy-to-use menu-driven interface for researchers not familiar with R:



Analysis software for EWAS

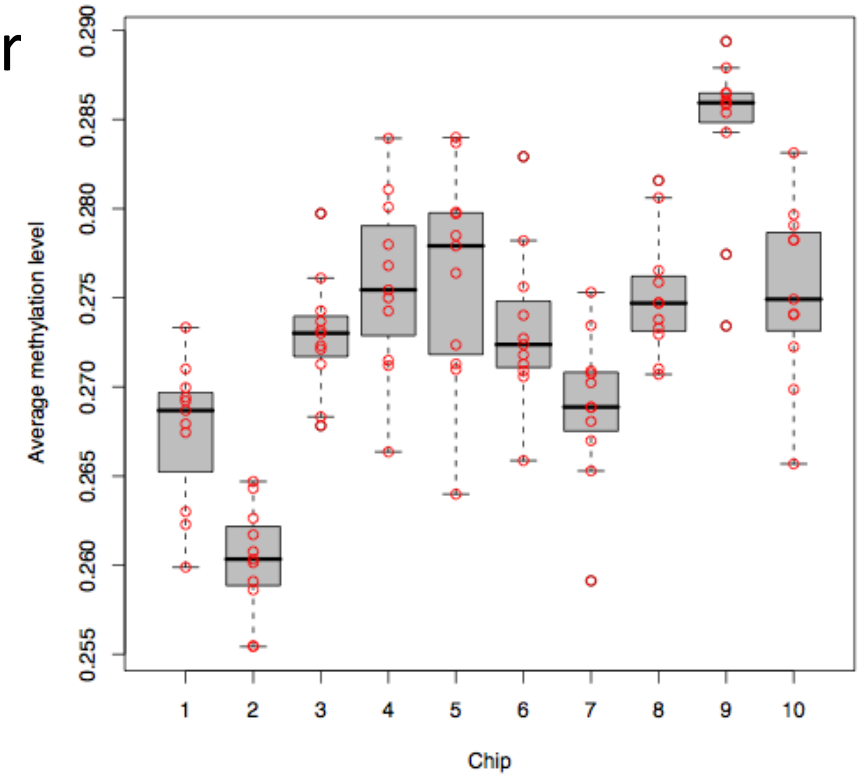
- Both CpGassoc and MethLAB perform fixed effects analyses rapidly by strategically partitioning the data
  - CpG sites with no missing observations are analyzed in a single matrix multiplication step
  - CpG sites with missing data are analyzed separately via a loop
  - Large datasets are partitioned further as necessary, based on a query of memory available in the environment

Data size	Fixed effects model with no missing data	Fixed effects model with 5.9% missing data	Mixed effects model with 5.9% missing data
27 k, $n = 200$	2	10	971
27 k, $n = 1000$	14	73	1409
450 k, $n = 200$	37	153	15 744
450 k, $n = 1000$	658	1621	22 993

Times in seconds, based on Intel Xeon 2.8 GHz, with 12 GB RAM. Model includes one covariate and either fixed or random chip effects.

1. Experiment, batch, and chip effects

- Measurement of methylation signals can vary due to differences in:
  - DNA quality/storage conditions (experiment/batch)
  - Bisulfite conversion (experiment/batch)
  - Handling of arrays (experiment, batch, and chip)
- These differences can occur even when data pass the typical QC measures
- Good study design and data analysis can minimize the impact of these differences

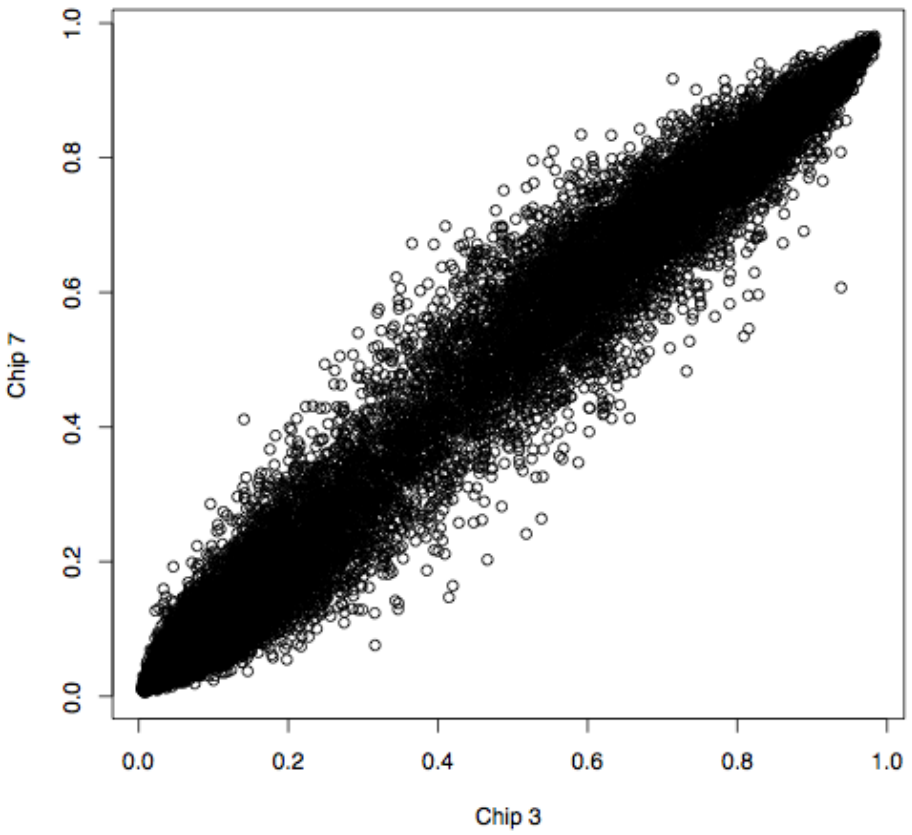


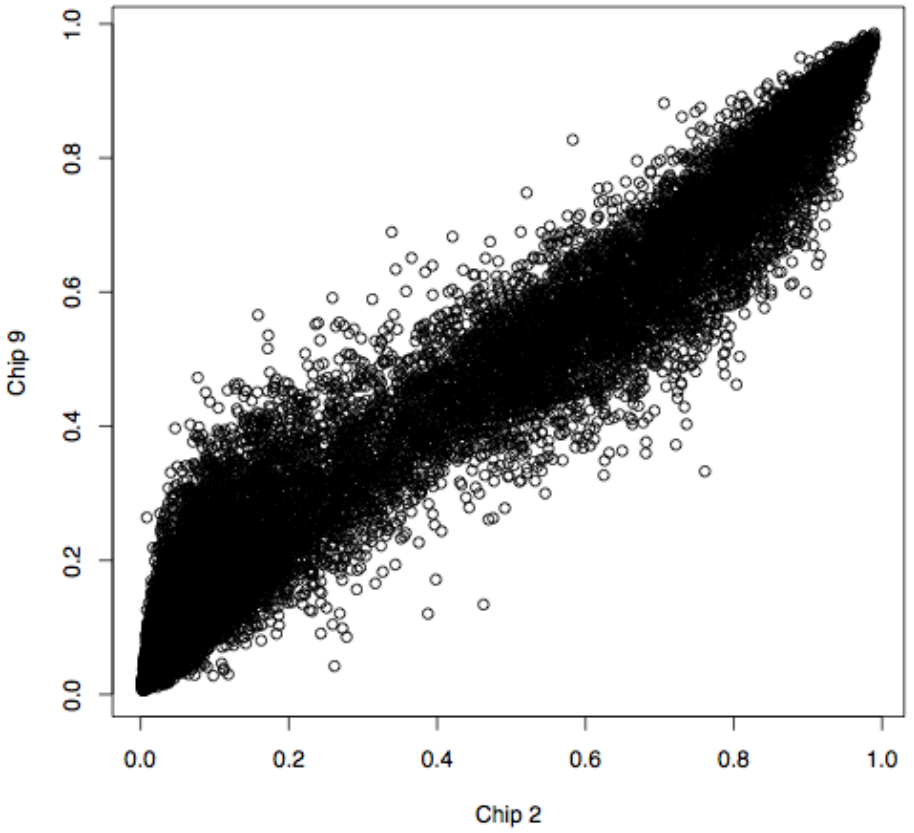
Design studies to avoid confounding due to chip effects

Randomize samples to chips to ensure phenotype of interest is evenly distributed

Include a technical control on every chip to compare performance of chips

Below: good and bad correspondence between technical controls across chips in Illumina 27K data



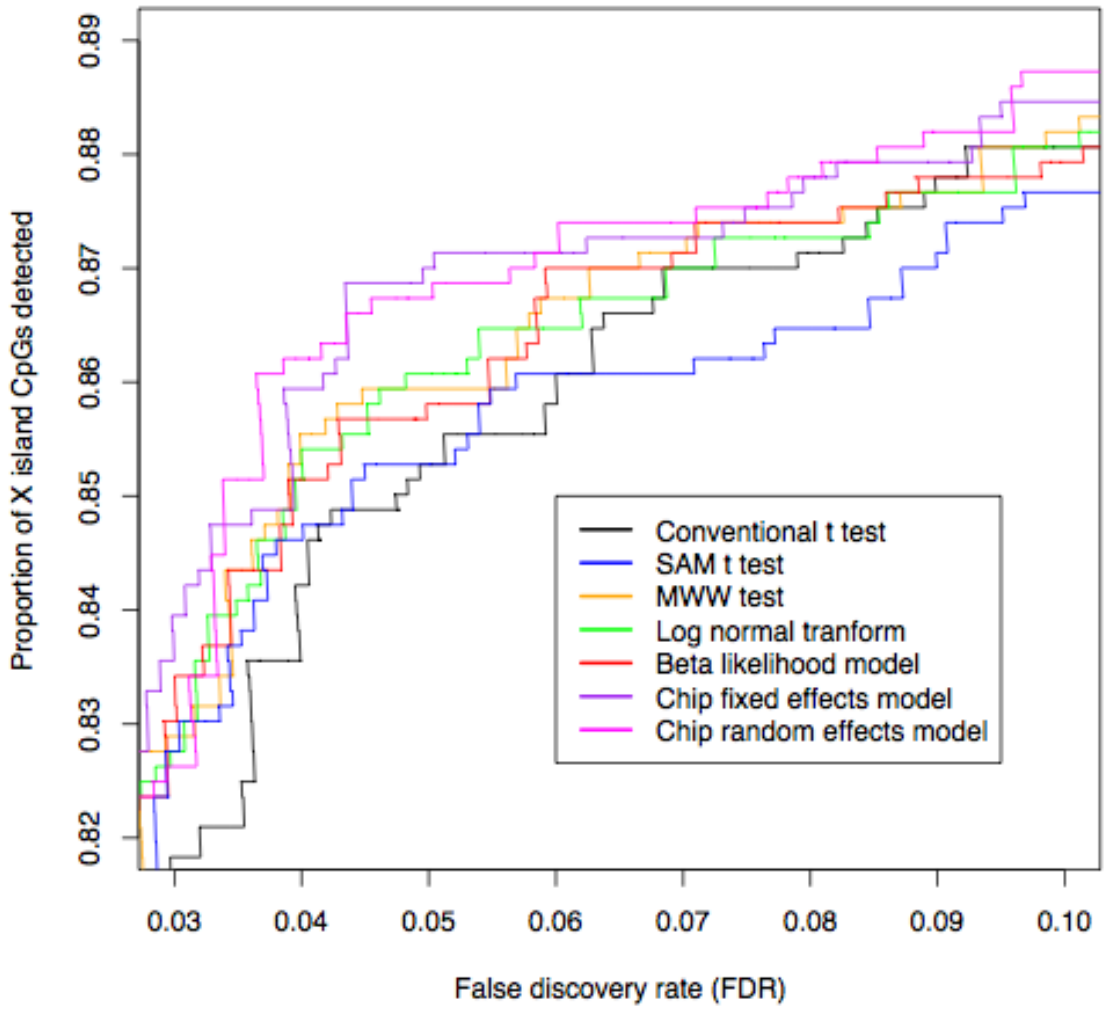


Adjustment for chip or batch

- We've already randomized (or stratified) individuals to chips to ensure that there is no relation between chip and our variables of interest
- Do we still need to adjust for chip?
- At this point, including a chip indicator as a covariate can only help us
  - For each chip in the experiment, 0-1 indicator variable indicating whether individual was on that chip
  - Including these covariates will account for chip effects that would otherwise be attributed to random noise
  - Accounting for this noise increases power to detect effects due to our variable of interest

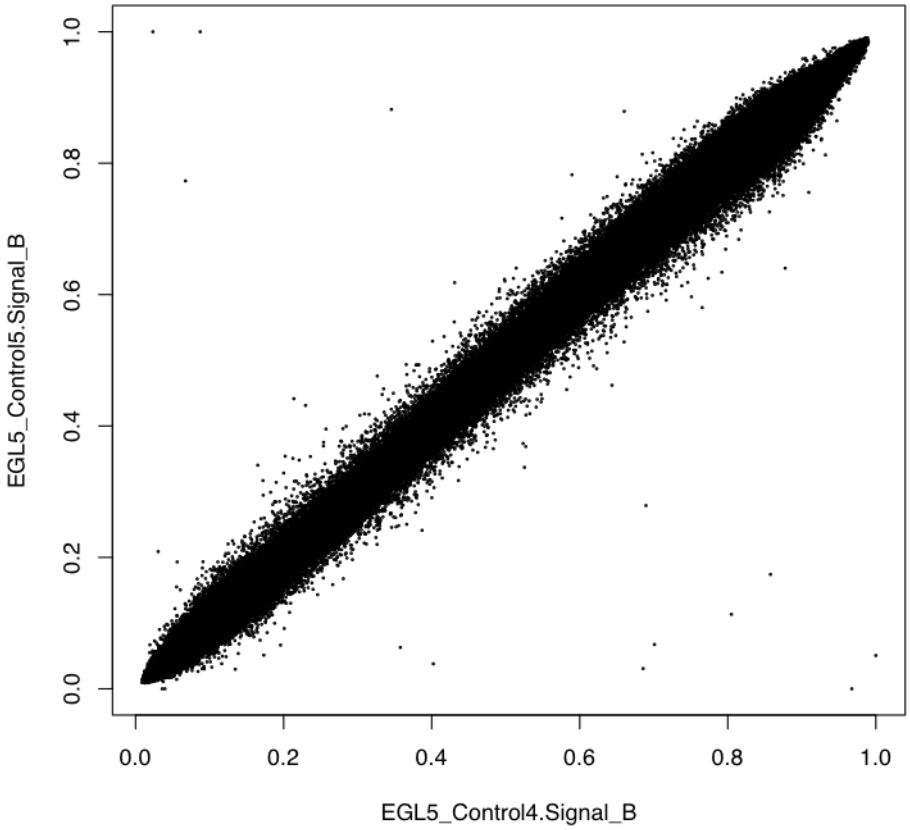
Adjusting for chip effects can increase power

- 67 women and 41 men randomly stratified to 27k chips with respect to sex
- Adjusting for chip effects still increased power to detect X CpGs
- We increased power through removal of noise



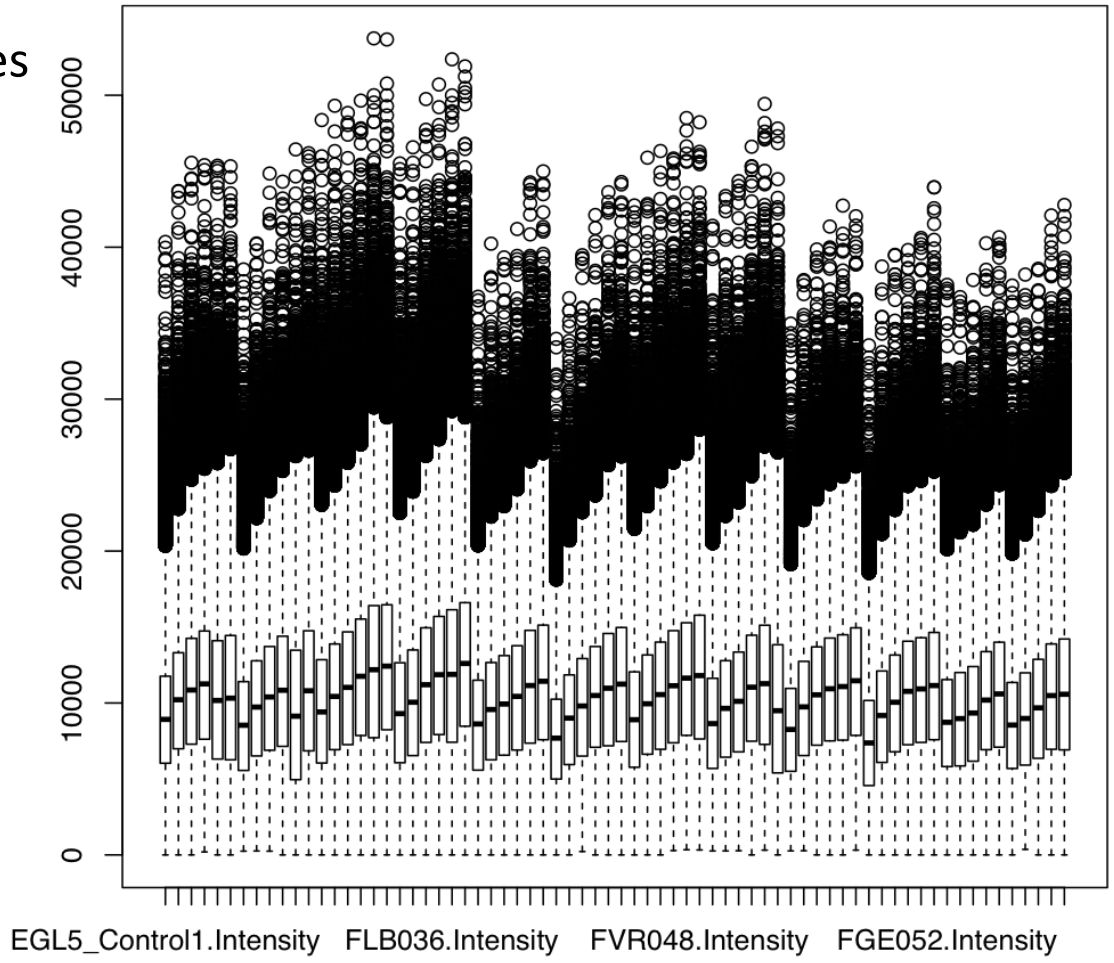
Are chip effects also an issue for 450k chips?

- Reproducibility seems to be quite high in general
- This is true regardless of any normalization
- Sometimes outliers exist even after stringent QC
- Solution: for smaller studies, duplicate or triplicate samples
- For large studies, include some duplicates (technical replicates)



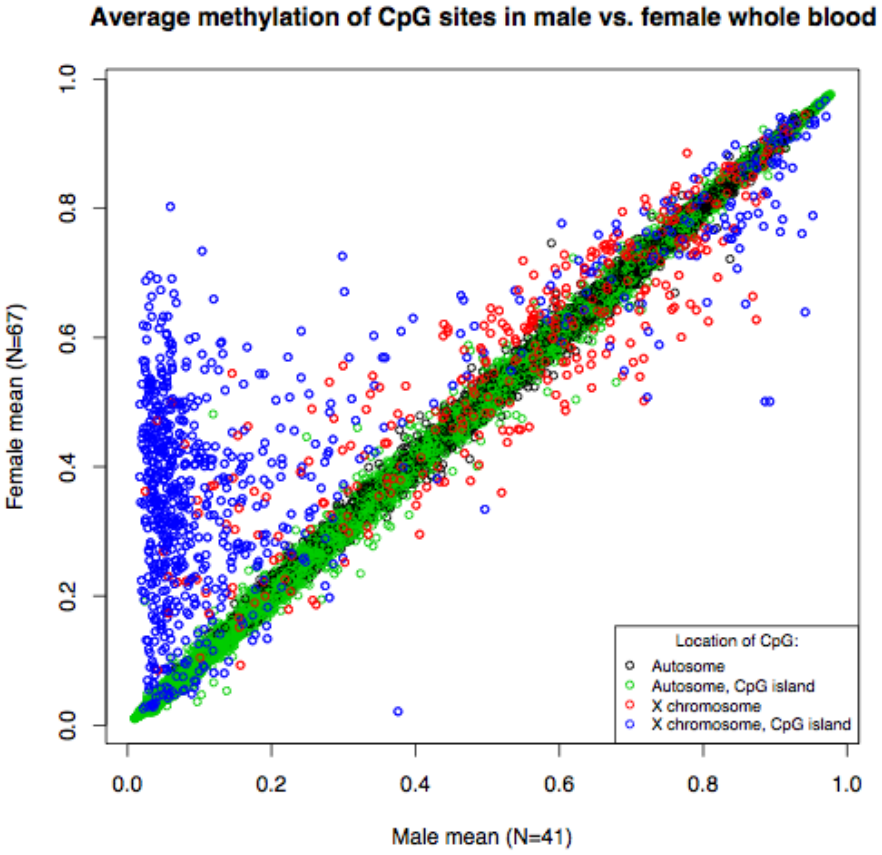
On 450k array, both chip and positional effects

- Each chip holds 12 samples
  - 6 rows
  - 2 columns
- Signal varies according to row on chip
  - “Gradient” effect
- Typical to include ‘row’ as an additional covariate in regressions



2. Sex

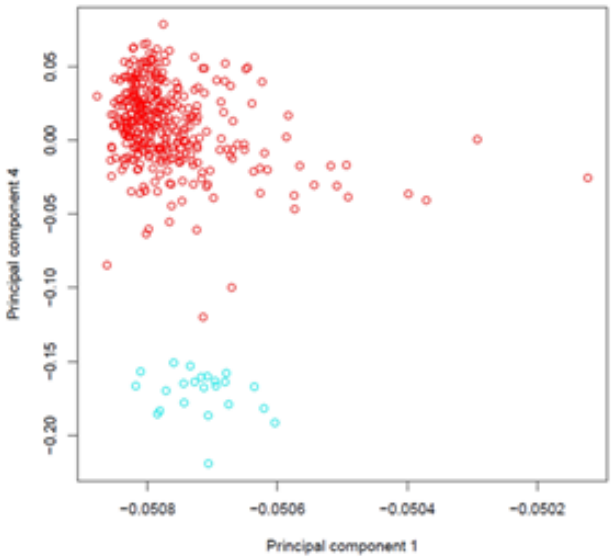
- Also common to include sex as a covariate
- Mainly need to do this for CpG sites on X chromosome
- However, other sites may vary by sex as well
  - Homologous probes
  - Sex-specific methylation?



3. Ancestry/race

- Why might DNA methylation vary by population?
  - SNP allele frequencies vary with population
  - Measured methylation can vary with genotype
    - A SNP right on a CpG site can abolish the site
    - A SNP in the primer region can influence measurement
    - Other forms of allele-specific methylation
  - Another possibility – differences in environment can influence methylation

Principal components 1 and 4 from CpG sites within 50 bp of SNPs



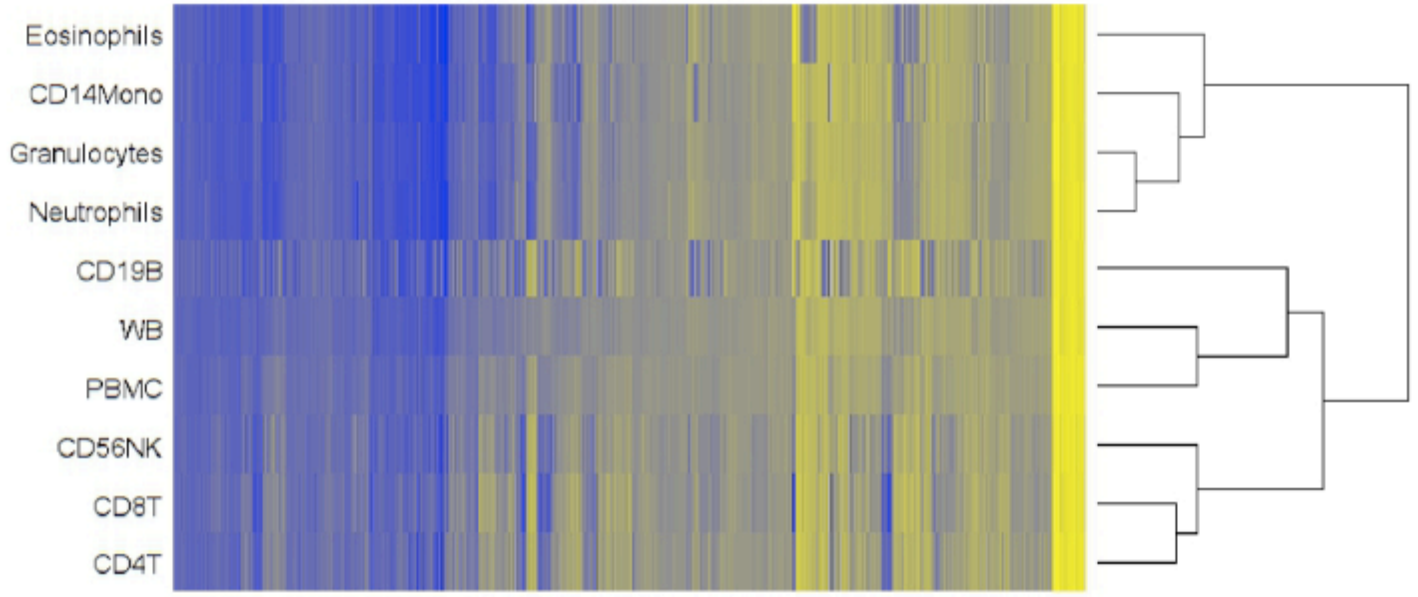
- As with genetic association studies, good to match and stratify by any available information on race/ethnicity
- Can also adjust for population stratification using standard methods such as principal component analysis with methylation data (Barfield et al., 2012)

4. Tissue and cell type

- Unlike genotype, epigenotype is tissue and cell-specific
  - Can potentially differ between any two cells
  - Highly variable between different tissues, cell types
- Choice of tissue to study may influence results
  - Depending on trait of interest, may make sense to study a specific tissue (e.g., brain, placenta)
  - Problem: easy to obtain blood, saliva, hard to obtain brains
- Even within a tissue, methylation may differ
  - Different regions of brain, placenta, etc.
  - DNA derived from blood comes from white blood cells, a mixture of neutrophils, lymphocytes, monocytes, eosinophils, basophils
  - Cell composition may vary across individual samples – yet another potential confounder!

Cell-type-specific methylation profiles

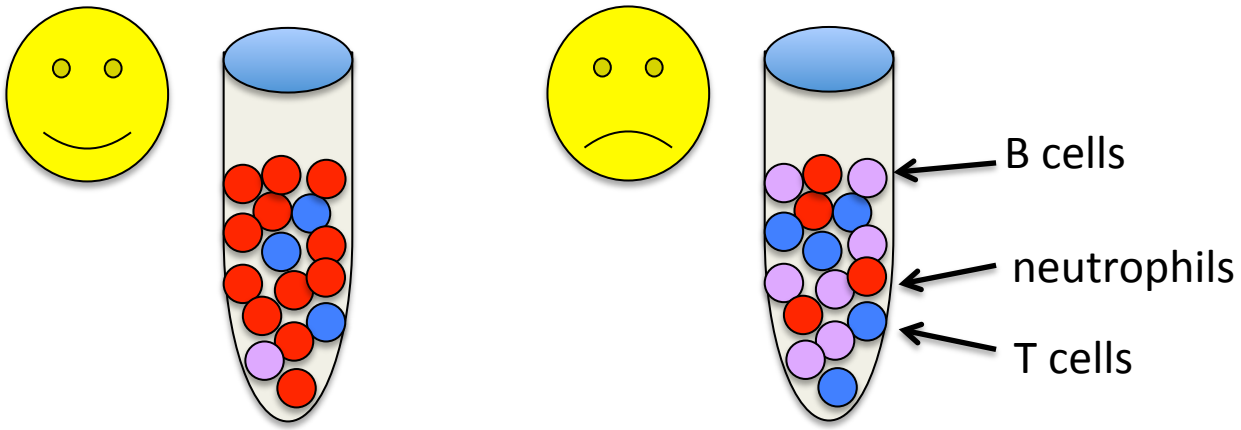
- Reinius et al. (*PLoS ONE*, 2012) examined genome-wide methylation profiles in homogenous cell populations
- Profiles for 343 genes implicated in immune-related GWAS



- 85% of CpG sites in these genes were significantly different between groups

Cell type as a possible confounder

- How do we address that methylation is measured on a population of heterogeneous cell types?
  - Example: Blood samples from two individuals:



  - Issue: are differences in methylation due to disease, or cell type composition?
  - Strategies: restrict to one cell type (difficult but possible), or measure and adjust for cell composition
  - If we have a complete blood count (CBC), we can include cell type proportions as covariates
  - What if we don't have one?

Regression approach to estimate cell type proportions

- A method by Houseman et al. (*BMC Bioinformatics*, 2012) is commonly adapted for this purpose[1]
- Start with reference sample of homogeneous cell types
- We observe methylation for the reference sample (β_{ref}) and for our target dataset (β_{data})
- The cell type identities / mixing proportions are known for the reference sample (Γ_{ref}) but not for our data (Γ_{data} = ?)
- Underlying model:

$$\beta_{ref} = \mathbf{B}_0 \Gamma_{ref} + \varepsilon_{ref}$$

$$\beta_{data} = \mathbf{B}_0 \Gamma_{data} + \varepsilon_{data}$$

- Step 1: linear regression on reference data to estimate B_0
- Step 2:

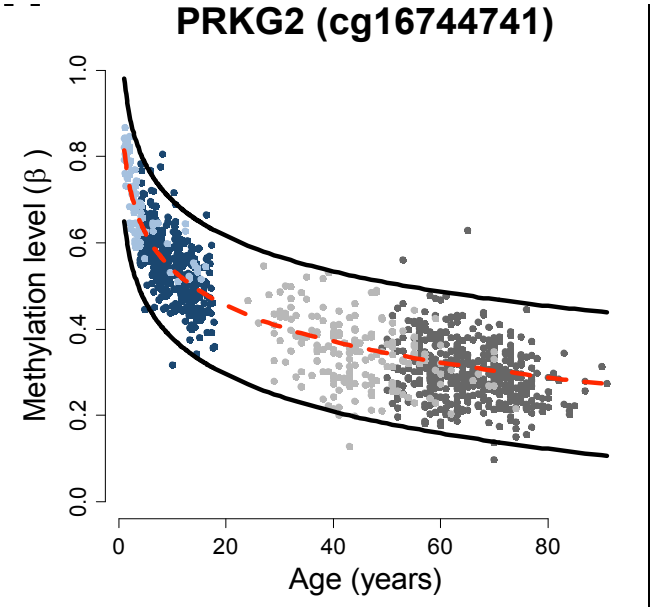
$$\hat{\Gamma}_{data} = \left(\hat{\mathbf{B}}_0^T \hat{\mathbf{B}}_0 \right)^{-1} \hat{\mathbf{B}}_0^T \beta_{data}$$

- Estimated proportions $\hat{\Gamma}_{data}$ can then be included as covariates in the methylation association test

[1] Now implemented for Illumina 450K data in *minfi* R package, with reference data from Reinius et al.

5. Age

- Have observed
  - Lifelong methylation trends for thousands of CpG sites across the genome [1]
  - Replicable methylation differences associated with gestational age in neonates [2]



[1] Alisch et al, *Genome Research* 2012
[2] Schroeder et al, *Epigenetics* 2011

Age as a confounder

- In genetic association studies, we don't worry about age
  - Genotype is static, so age at sample collection is irrelevant
  - Often makes sense to use older controls to ensure that they're truly controls
- In methylation studies, age is a confounder
  - Methylation is known to change with age; crucial to development
  - Analogous to population stratification in genetic association studies
- As with chip effects, solution is two-pronged:
  - Cases and controls should be age-matched to avoid confounding
  - If not possible, try to keep them in a similar range
  - Additionally, adjustment for age as a covariate will increase power (by removing effects that would otherwise be considered noise)

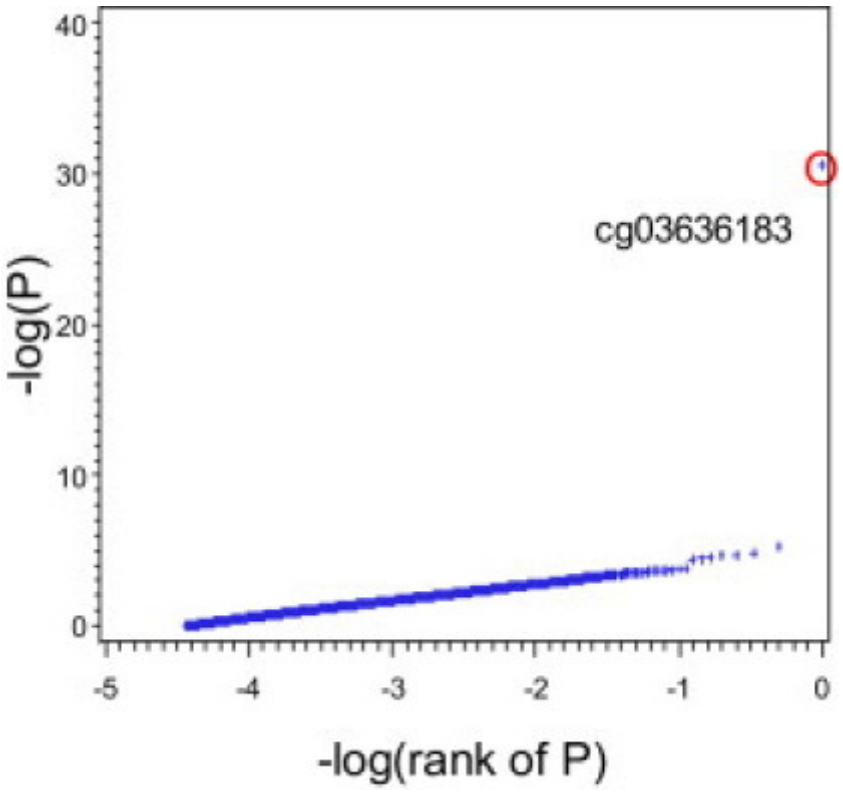
Finally! Performing the EWAS

- For each CpG site, fit a linear regression that includes relevant covariates:

$$\beta_i = \alpha_0 + \alpha_1 \cdot X_i + \alpha_2 \cdot \mathrm{cov}_1 + \alpha_3 \cdot \mathrm{cov}_2 + \varepsilon_i$$

- As with a GWAS, this analysis will generate p-values for thousands of sites

- Example: a CpG site in *F2RL3* highly associated with tobacco smoking
- Breitling et al. (AJHG, 2011)
- Replicated many times



Multiple test correction

- As with genetic association studies, need to adjust for the many tests performed
- The smoking CpG site was a smoking gun
  - Easily passed Bonferroni correction
  - No other sites were significant in Breitling study
- In other studies, we may expect to see large numbers of mildly associated sites
  - Bonferroni adjustment is most appropriate if we expect to see no more than a handful of associated genes (ie, in a GWAS setting)
  - If we expect to see large sets of CpG sites correlated with trait variable (e.g., entire gene networks), a less conservative criterion may be appropriate

Family-wise error rate (FWER)

FWER = probability at least one true hypothesis is rejected

= P(V > 0)

Frequency distribution of hypotheses rejected and not rejected

	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m-m_0
Total	m-R	R	m

FWER is controlled by the **Bonferroni** adjustment

Is FWER always what we're interested in?

Storey, 2003:

The FWER offers an extremely strict criterion which is not always appropriate. It is possible for a multiple hypothesis testing situation to exist in which one is more concerned about the rate of false positives among all rejected hypotheses rather than the probability of making one or more Type I errors. We have seen a recent increase in the size of data sets available. It is now often up to the statistician to find as many interesting features in a data set as possible rather than test a very specific hypothesis on one item. For example, one is more frequently faced with the daunting task of estimating or performing hypothesis tests on thousands of parameters simultaneously. In this kind of situation, one is more interested in the total number of false positives compared to the total number of significant items, rather than making one or more Type I errors.

The Annals of Statistics
2003, Vol. 31, No. 6, 2013–2035

False discovery rate (FDR)

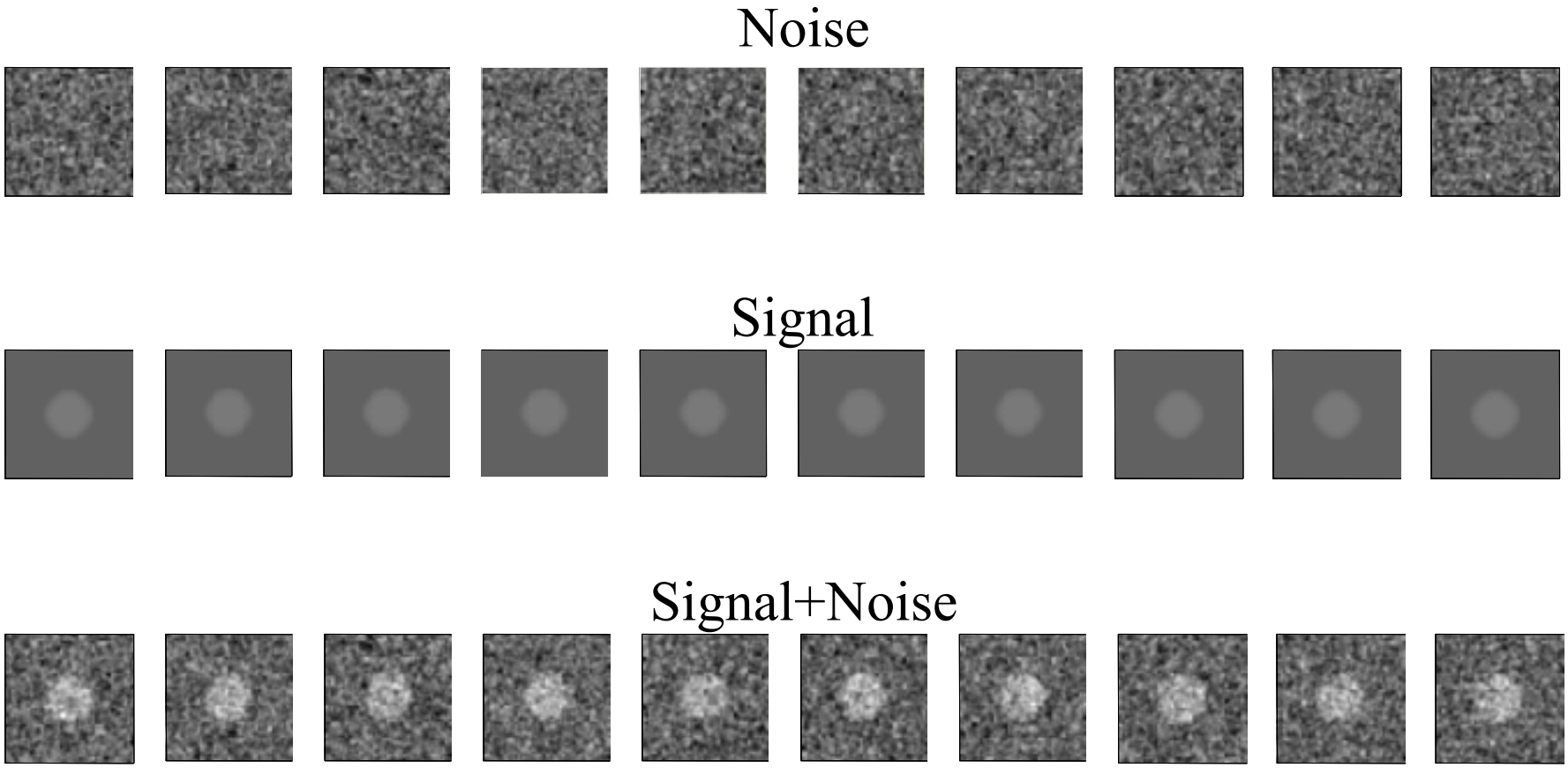
FDR = Expected proportion of rejected hypothesis that were actually true = E(V/R) for R > 0

If R = 0 (ie, no hypotheses actually rejected), then FDR = 0

Frequency distribution of hypotheses rejected and not rejected

	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m-m_0
Total	m-R	R	m

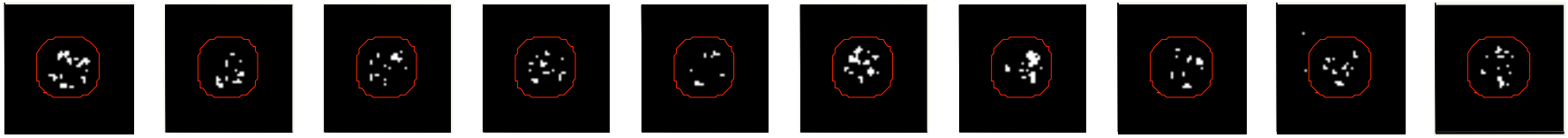
False Discovery Rate Illustration:



Source: Tom Nichols, http://www.sph.umich.edu/~nichols/FDR/SMU2004.ppt

False Discovery Rate Illustration:

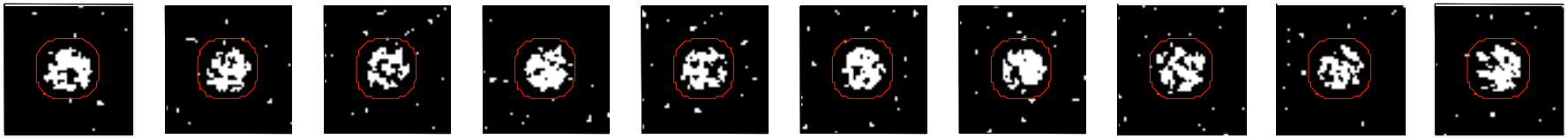
Control of Familywise Error Rate at 10%



FWE

Occurrence of Familywise Error

Control of False Discovery Rate at 10%



6.7% 10.4% 14.9% 9.3% 16.2% 13.8% 14.0% 10.5% 12.2% 8.7%

Percentage of Activated Pixels that are False Positives

Source: Tom Nichols, http://www.sph.umich.edu/~nichols/FDR/SMU2004.ppt

FDR vs. FWER

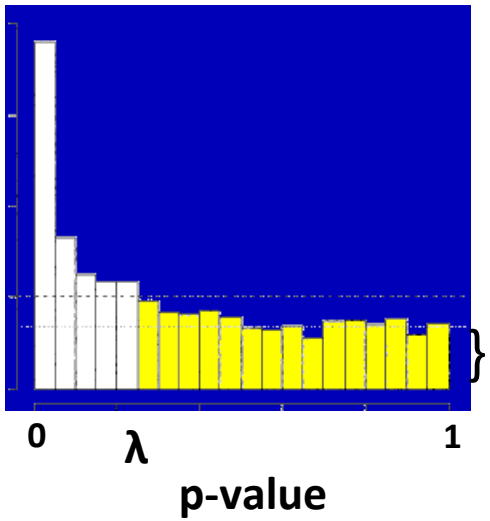
	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m-m_0$
Total	m-R	R	m

- FDR = E(V/R) = expected proportion of rejected hypotheses that are actually true
- FWER = P(V>0) = probability of rejecting *any* true hypotheses

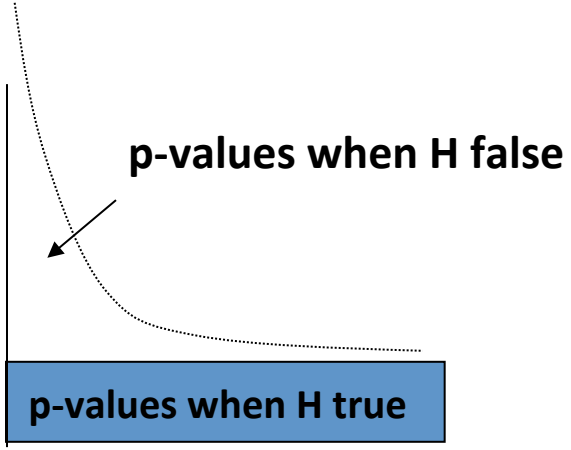
Controlling the FDR: Storey-Tibshirani

- Choose a significance threshold λ so that:
  - For tests with p>λ, distribution of p-values appears uniform
  - For tests with p<λ, q are false positives and 1-q are true positives (where q is the FDR we are comfortable accepting)

Distribution of observed p-values



=



Source: Storey and Tibshirani (2003)

Controlling the FDR: Benjamini-Hochberg

- Benjamini-Hochberg (B-H) procedure controls FDR
  - Order p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
  - Let $k = \max\{i\}$ such that $p_{(i)} \leq \frac{i}{m} q$ where q is the level at which we desire to control FDR ($E(V/R) < q$)
  - Reject $H_{(1)} \ldots H_{(k)}$
  - Will also control FWER when all null hypotheses are true

Source: Benjamini and Hochberg (1995)

Follow-up of significant results: testing for biological enrichment

- Depending on the trait or exposure studied, you may end up with a set of significant results (FDR- or Bonferroni-significant)
  - Studies of methylation and complex disease in a single tissue: usually only a handful of significant hits, if any
  - Other types of studies may lead to a large set of significant CpG sites:
    - Inter-tissue studies (eg tumor vs. normal)
    - Studies of methylation and aging
    - Studies of extreme environmental exposures
- To make sense of large sets of significant CpG sites, we can:
  - Test for enrichment of gene networks
  - Test for enrichment of biological features (eg CpG islands, DNAse hypersensitivity)

Enrichment of biological features

- Scenario: We test 485,096 CpG sites for association with age.
- 14,102 sites are significant after Bonferroni adjustment!
  - $p < .05/485{,}096 = 1.03 \times 10^{-7}$
- Can examine CpG island status of significant sites

Table 2: Summary of age-associated CpG sites

Direction of association with age	N	% on CpG island	% on CpG shore	% on CpG shelf	% not near CpG island
Positively associated	9,503	74.1	17.3	1.7	6.9
Negatively associated	4,599	4.7	36.9	12.0	46.3
Not associated	470,994	30.3	23.1	9.8	36.8

- It looks like sites showing increasing methylation with age are more likely to be on islands than unassociated sites (*enriched*)
- Similarly, positively associated sites appear *depleted* for shelves
- How can we test this formally?

A quick do-it-yourself enrichment test

Table 2: Summary of age-associated CpG sites					
Direction of association with age	N	% on CpG island	% on CpG shore	% on CpG shelf	% not near CpG island
Positively associated	9,503	74.1	17.3	1.7	6.9
Negatively associated	4,599	4.7	36.9	12.0	46.3
Not associated	470,994	30.3	23.1	9.8	36.8

- Put observed counts into 2x2 table:

OBSERVED	On island	Not on island	Total
Pos. associated	7,042	2,461	9,503
Unassociated	142,711	328,233	470,994
Total	150,113	330,384	480,497

- 150,113/480,497 = .3124 → on average, 31.24% on islands
  - Thus, we would randomly expect .3124 × 9,503 = 2,969 of positively associated CpG sites to be on islands.

EXPECTED	On island	Not on island	Total
Pos. associated	2,969	6,534	9,503
Unassociated	147,144	323,800	470,994
Total	150,113	330,384	480,497

A quick do-it-yourself enrichment test

OBSERVED	On island	Not on island	Total
Pos. associated	7,402	2,101	9,503
Unassociated	142,711	328,282	470,994
Total	150,113	330,384	480,497

EXPECTED	On island	Not on island	Total
Pos. associated	2,969	6,534	9,503
Unassociated	147,144	324,044	470,994
Total	150,113	330,384	480,497

- So what can we do with our tables of observed & expected counts?
- Test for significant deviation of observed (O) from expected (E):
  - Chi-squared test: $$X^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}\frac{\left(O_{ik}-E_{ik}\right)^2}{E_{ik}} \sim \chi^2_{(I-1)(K-1)}$$
  - Fisher's exact test (the chi-squared test is just an approximation to this test).

Review of Fisher's exact test

	Men	Women	*Row Total*
Dieting	***a***	***b***	*a + b*
Non-dieting	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

- The odds ratio from this comparison is OR = ad/bc
- To get a p-value, we would
  - 1) compute *p* above for all combinations of a, b, c, and d that led to an OR greater than or equal to the original OR
  - 2) sum all of these *p*s.

A quick do-it-yourself enrichment test

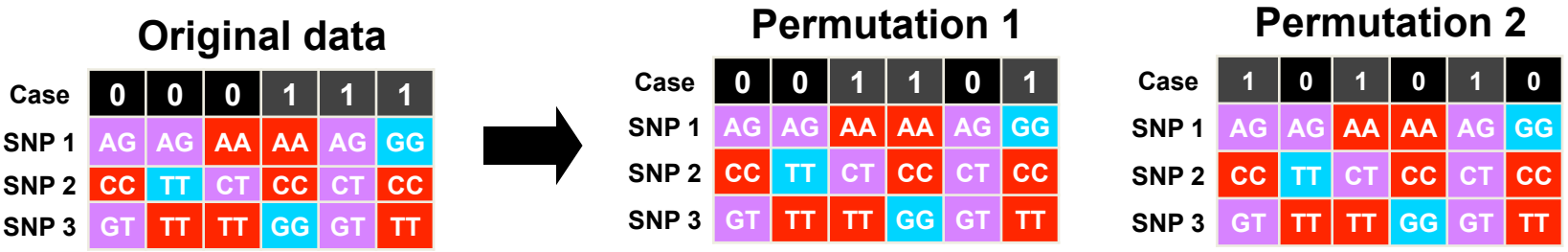
OBSERVED	On island	Not on island	Total
Pos. associated	7,402	2,101	9,503
Unassociated	142,711	328,282	470,994
Total	150,113	330,384	480,497

EXPECTED	On island	Not on island	Total
Pos. associated	2,969	6,534	9,503
Unassociated	147,144	324,044	470,994
Total	150,113	330,384	480,497

- Both chi-squared and Fisher's tests test whether *association status* is independent of *island status*
- Both depend on the assumption that each observation is independent
- Here, our observations are CpG sites. Are these independent of one another, or are they likely to be correlated in some way?
- This is useful for quick analyses, but for publication it's worth following up with permutation tests (will be covered in lab exercise.)

Permutation tests

- Permutation allows estimation of unbiased *p*-values
  - For each permutation, randomly shuffle case/control status
    - Random assignment simulates null hypothesis of no association
    - Permute in a way that leaves correlation structure intact
  - Re-perform association analysis for each permutation
  - Example from a SNP association study



  - $p_{permutation} = \frac{\text{\# times result as extreme as original result}}{\text{\# permutations}}$

Summary: analysis of methylation data

- We talked about one common method for measuring methylation on a large scale
  - Illumina bead array approach
  - Relies on bisulfite treatment of DNA
  - Estimate proportion of DNA methylated at single-CpG-site resolution
- DNA collected for GWAS is increasingly utilized in methylation studies, but there are important differences to consider
  - GWAS cases and controls generally not age-matched
  - Tissue type not important for GWAS, can be important for methylation
  - Often reasonable to use a less stringent multiple test correction (FDR)
- Other issues remain important for both types of studies
  - Validation of key results through
    - permutation testing
    - biological validation (generally sequencing of regions surrounding sites of interest)
  - Replication in additional samples