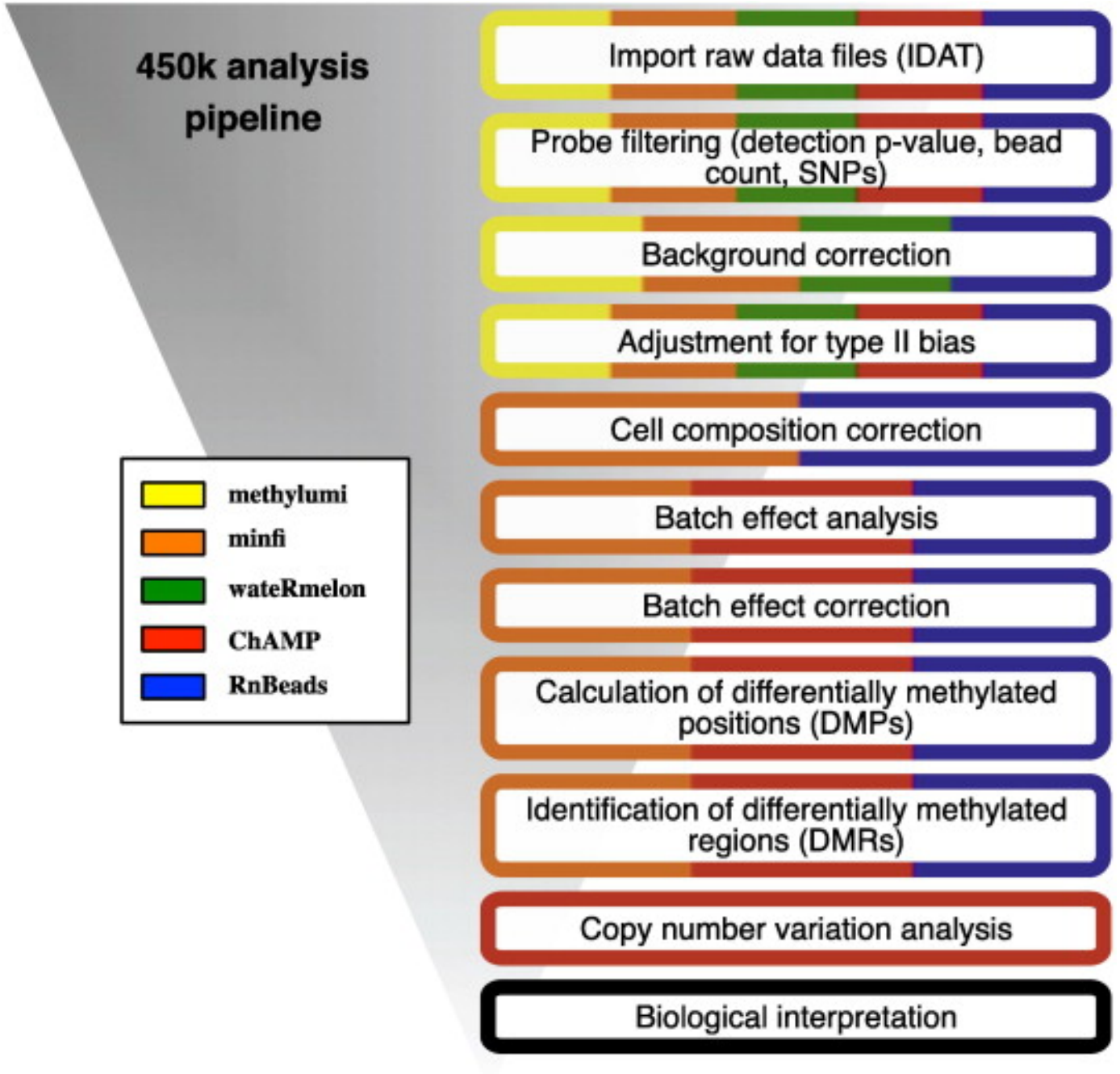
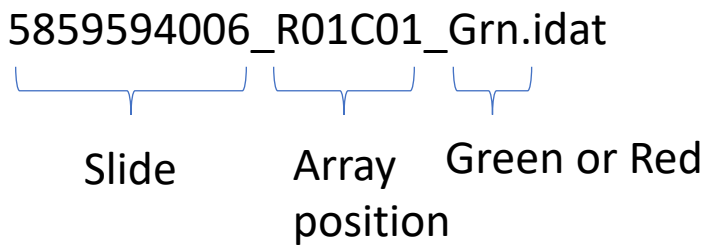
# Methylation Array Workflow

# minfi

- Tools for analyzing Illumina arrays
- Provides tools for many of the steps presented here.



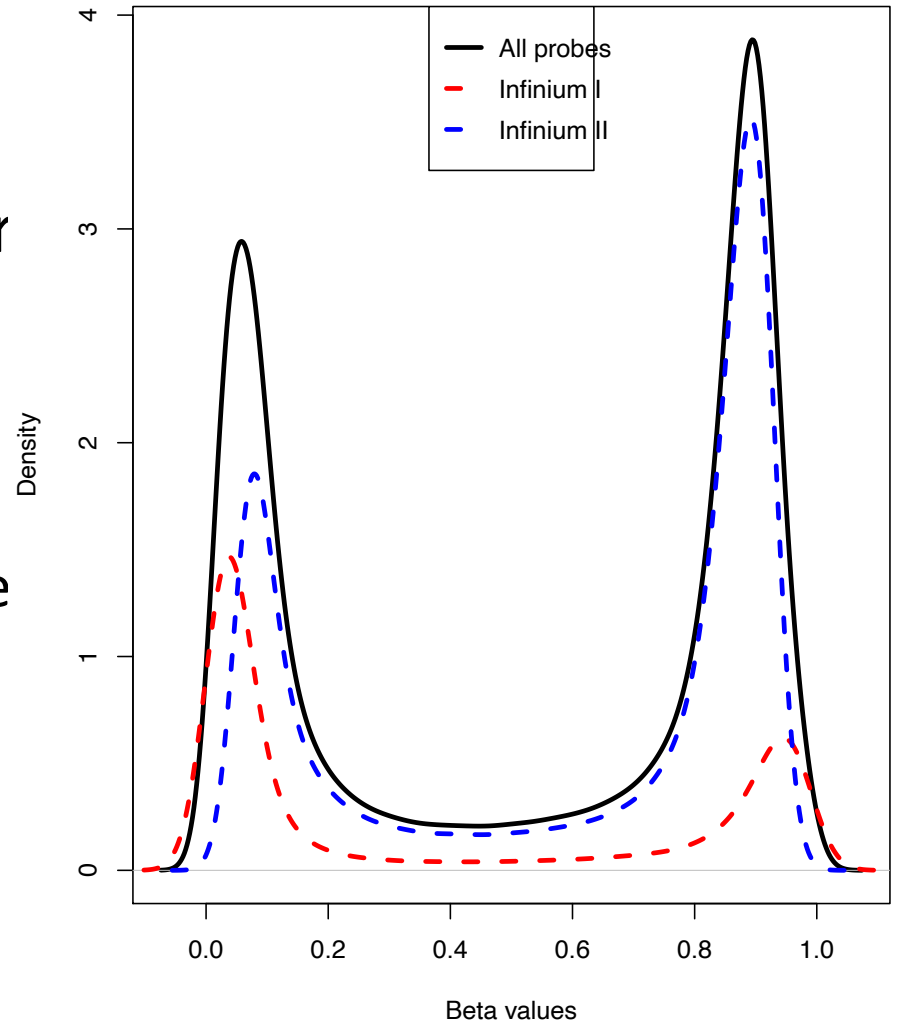
# Import Data




- IDAT files; slide scanner output
- Needs a SampleSheet, usually accompanies array data (or can be made manually)
- Raw intensities -> *RGChannelSet*
- Needs to be converted to *MethylSet* for initial QC

# QC + Filtering

- Aim: find outliers/batch effects and artifacts and try to remove or account for them
- Several metrics:
  - Plot distributions of the Beta values
  - Quality of probes: average detection p-value
  - Internal quality control probes



# QC + Filtering

- Aim: find outliers/batch effects and artifacts and try to remove or account for them
- Several metrics:
  - Plot distributions of the Beta values
  - Quality of probes: average detection p-value
  - Internal quality control probes
  - Remove probes with known SNPs
  - MDS/PCA plot

- STAINING CONTROLS
- BISULFITE CONVERSION CONTROLS
- EXTENSION CONTROLS
- SPECIFICITY CONTROLS
- HYBRIDIZATION CONTROLS
- TARGET REMOVAL CONTROLS
- NON-POLYMORPHIC CONTROLS
- NEGATIVE CONTROLS

# Normalization

- Within and across array normalization

**Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies**

Paul Yousefi, Karen Huen, Raul Aguilar Schall, Anna Decker, Emon Elboudwarej, Hong Quach, ...show all

**Between-array normalization for**
**a**

*nd Hermann Brenner [1, 2]*

*ogy and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany,*

Functional normalization of 450k methylation array data improves replication in large cancer studies

Jean-Philippe Fortin[1], Aurélie Labbe[2,3,4], Mathieu Lemire[5], Brent W Zanke[6], Thomas J Hudson[5,7], Elana J Fertig[8], Celia MT Greenwood[2,9,10] and Kasper D Hansen[1,11*]

**A systematic assessmen**
**approaches for the Infin**
**platform**

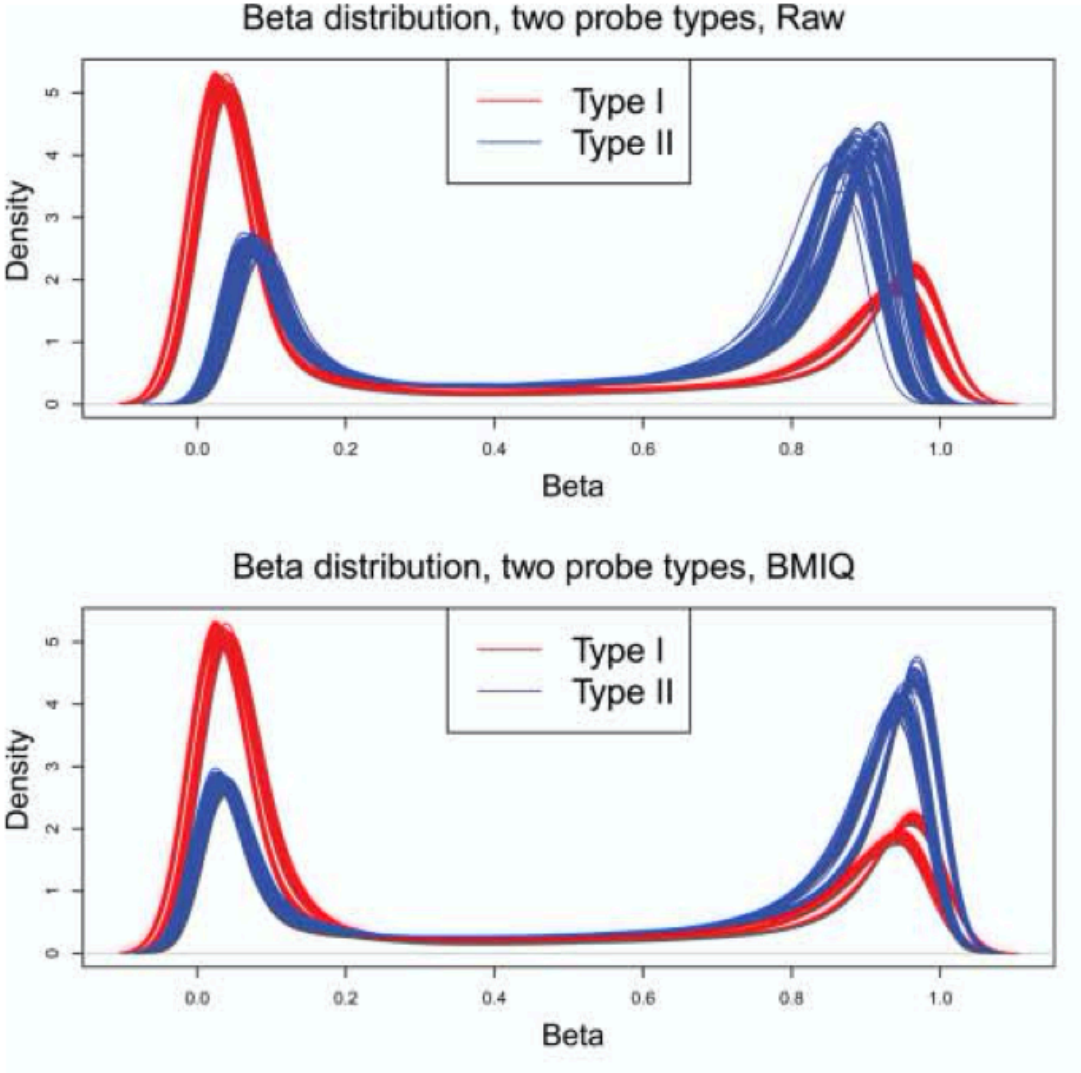
**Michael C Wu, Bonnie R Joubert, Pei-fen Kuan, Siri E Håberg, Wenche Nystad, Shyamal D Peddada & Stephanie J London**

**on methods for**
**ethylation data using whole-**
**e sequencing data**

Nadia Boutaoui, Glorisa Canino, Jianhua Luo, ...show all

# Normalization

- Within and across array normalization
- Crucial step;
  aims to make distribution more comparable

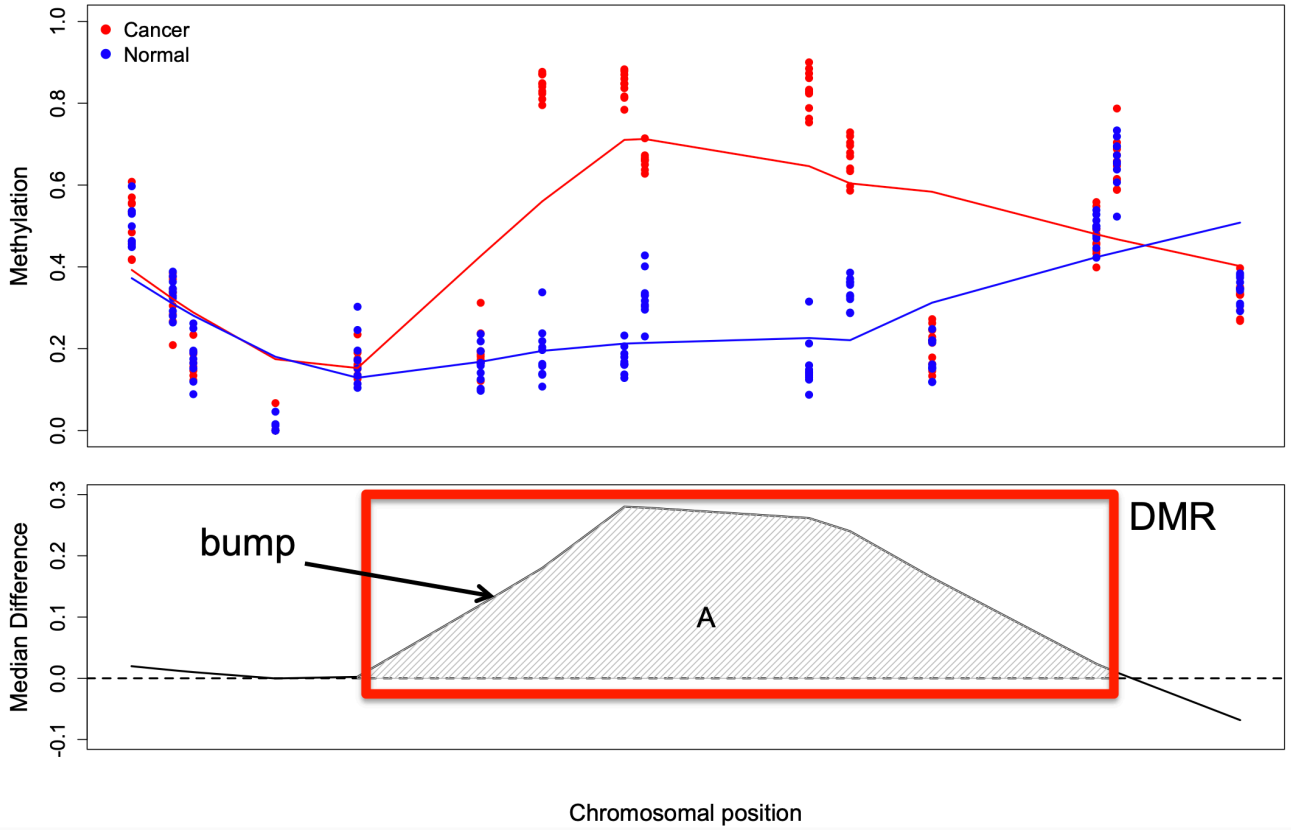


# Differential Methylation

- Identification of systematic differences in methylation between groups of samples (i.e., case vs control, smokers vs non-smokers, ...)
- Countless ways to approach this, depending on:
  - Question(s) being asked
  - Available information on potential confounders
  - Nature/structure of the data (repeat measurements, …)
- Some possible approaches include:
  - T-tests and ANOVA models
  - Wilcoxon rank-sum and Kruskal Wallis tests
  - Linear, logistic and Cox regression
  - Mixed effects models
  - Surrogate Variable Analysis (SVA)
- Use M-values: $Mvalue{=}log2(M/U)$
  - More homoscedastic

# Differential Methylation

- Single CpG can be useful (DMP), but often regions or block of CpGs (DMR)
- How to define region?
  - Sliding window
  - Heuristic cutoffs/Smoothing
  - Functional units



# Gene Set Enrichment

- Long list of DMP or DMR… What does it mean?
- Gene expression -> GO analysis
- Not so straightforward for methylation data!
  - CpG link to genes unclear
  - Directionality?
  - Extreme bias: number of CpG per gene differs

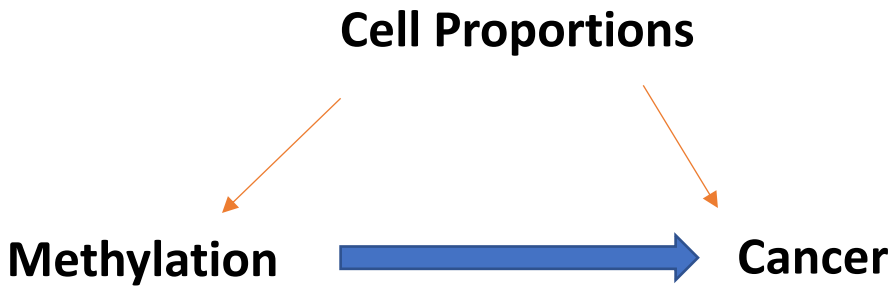
**Gene-set analysis is severely biased when applied to genome-wide methylation data**

Paul Geeleher[1,2], Lori Hartnett[3], Laurance J. Egan[3], Aaron Golden[4], Raja Affendi Raja Ali[3] and Cathal Seoighe[2,*]

- missMethyl, methylGSA, BioMethyl

# Cell Type Deconvolution

- Estimates the relative proportion of pure cell types within a sample
- *Minfi*: RGChannelSet from a DNA methylation study of blood, and return the relative proportions of CD4+ and CD8+ T-cells, natural killer cells, monocytes, granulocytes, and b-cells in each sample.
- Most cohort studies currently analyse data from blood samples: can be used to correct for cell type heterogeneity



# Datasets

- Small toy data
- IDAT files
- 10 samples in total: there are 4 different sorted T-cell types , collected from 3 different individuals :
  - Naïve
  - Treg
  - act_naive
  - act_Treg
- An additional birth sample is included from another study (GSE51180) to illustrate approaches for identifying and excluding poor quality samples.