Principal Component of Explained Variance

An efficient and optimal dimension reduction framework for association studies

Maxime Turgeon

PhD Candidate, McGill University

November 26th, 2015

Collaborators

This is joint work with:

- Karim Oualkacha (UQAM)
- Antonio Ciampi (McGill)
- Golsa Dehghan (Masters student, McGill)
- Brent Zanke (Ottawa Hospital Research Institute)
- Celia Greenwood (McGill)
- Aurélie Labbe (McGill)

Multidimensional phenotypes

Multidimensional phenotypes

- There are at least two ways in which multi-dimensional phenotypes are relevant:

Multidimensional phenotypes

- There are at least two ways in which multi-dimensional phenotypes are relevant:
    1. In complex diseases, one may be interested in studying the association between covariates and intermediate phenotypes, instead of the association with the disease status.

Multidimensional phenotypes

- There are at least two ways in which multi-dimensional phenotypes are relevant:
  1. In complex diseases, one may be interested in studying the association between covariates and intermediate phenotypes, instead of the association with the disease status.
  2. One may also be interested in the joint analysis of correlated phenotypes, to account for pleiotropy for example.

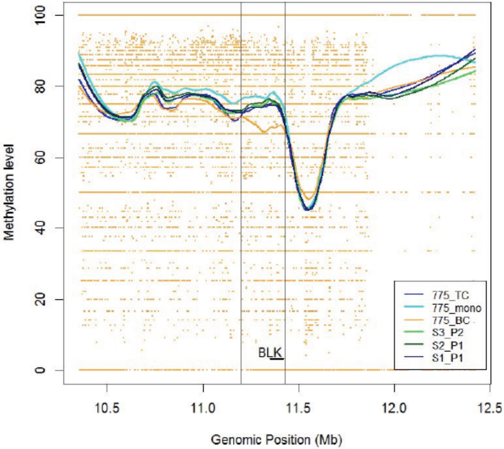
Multidimensional phenotypes

- There are at least two ways in which multi-dimensional phenotypes are relevant:
  1. In complex diseases, one may be interested in studying the association between covariates and intermediate phenotypes, instead of the association with the disease status.
  2. One may also be interested in the joint analysis of correlated phenotypes, to account for pleiotropy for example.
- In this setting, we want a statistical approach which can take the correlation into account and also reduce the overall dimension.

Motivating example

Motivating example

B-Lymphoid Tyrosine Kinase (BLK) gene is known to be differentially methylated with respect to blood cell types.



Motivating example

- The data consist of 40 cell-separated whole-blood samples (T cells, B cells, monocytes), for which methylation levels were measured at 24,000 CpG sites using bisulfite sequencing.

Motivating example

- The data consist of 40 cell-separated whole-blood samples (T cells, B cells, monocytes), for which methylation levels were measured at 24,000 CpG sites using bisulfite sequencing.
- The figure above was obtained using smoothing techniques: the methylation levels for a particular cell-type is smoothed across the 24,000 loci.

Motivating example

- The data consist of 40 cell-separated whole-blood samples (T cells, B cells, monocytes), for which methylation levels were measured at 24,000 CpG sites using bisulfite sequencing.
- The figure above was obtained using smoothing techniques: the methylation levels for a particular cell-type is smoothed across the 24,000 loci.
- Can we study this region using a dimension reduction approach, for example Principal Component Analysis?

PCA

PCA

- First, we will restrict ourselves to the 6,000 CpG most variable sites.

PCA

- First, we will restrict ourselves to the 6,000 CpG most variable sites.
- Then we can compute the first Principal Component PC_1.

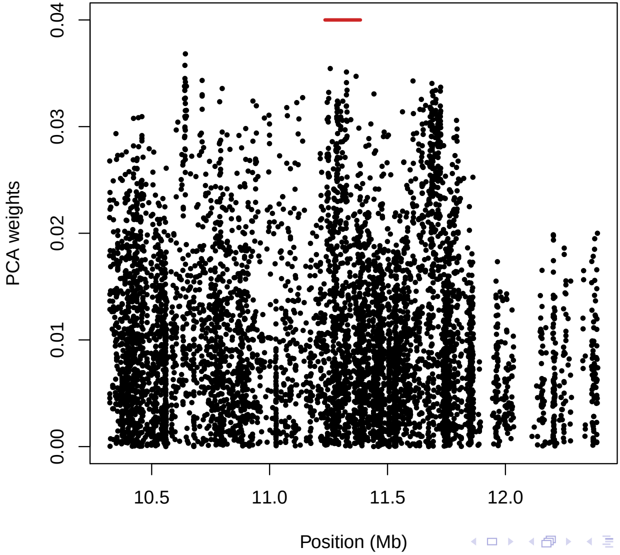
PCA

- First, we will restrict ourselves to the 6,000 CpG most variable sites.
- Then we can compute the first Principal Component PC_1.
- PC_1 can then be used in a linear regression model to test for association with a covariate. In our setting, we will focus on a binary covariate: $X_i = 1$ if sample i comes from a B cell and $X_i = 0$ otherwise.

PCA

- First, we will restrict ourselves to the 6,000 CpG most variable sites.
- Then we can compute the first Principal Component PC_1.
- PC_1 can then be used in a linear regression model to test for association with a covariate. In our setting, we will focus on a binary covariate: $X_i = 1$ if sample i comes from a B cell and $X_i = 0$ otherwise.

If we do this on the same data that was used for the previous figure, we get a p-value of 0.7415...


PCA weights
0.00
0.01
0.02
0.03
0.04
10.5
11.0
11.5
12.0
Position (Mb)

PCA

The problem with PCA in this setting is that it “picks up” the most variable direction. But most of the variance is explained by other factors than cell type, making it inappropriate for this particular analysis.

Most important slide

Most important slide

- In this talk, I will present a similar dimension reduction approach that can uncover the **direction most associated with a set of covariates**.

Most important slide

- In this talk, I will present a similar dimension reduction approach that can uncover the **direction most associated with a set of covariates**.
- I will explain how it can be used in **association studies**.

Most important slide

- In this talk, I will present a similar dimension reduction approach that can uncover the **direction most associated with a set of covariates**.
- I will explain how it can be used in **association studies**.
- I will show how it can be efficiently computed in a **high-dimensional setting** (when the number of phenotypes is larger than the sample size).

Most important slide

- In this talk, I will present a similar dimension reduction approach that can uncover the **direction most associated with a set of covariates**.
- I will explain how it can be used in **association studies**.
- I will show how it can be efficiently computed in a **high-dimensional setting** (when the number of phenotypes is larger than the sample size).

All this is implemented in an `R` package called `pcev`, currently hosted on Celia's Lab's GitHub account (`https://github.com/GreenwoodLab/pcev`) but soon to be sent to `CRAN`.

PCEV: Statistical model

PCEV: Statistical model

Let $\mathbf{Y}$ be a multidimensional vector of phenotype values of dimension p and X, a vector of covariates.

PCEV: Statistical model

Let $\mathbf{Y}$ be a multidimensional vector of phenotype values of dimension p and X, a vector of covariates.
We assume a linear relationship:

$$\mathbf{Y} = BX + E.$$

PCEV: Statistical model

Let $\mathbf{Y}$ be a multidimensional vector of phenotype values of dimension p and X, a vector of covariates.
We assume a linear relationship:

$$\mathbf{Y} = BX + E.$$

The total variance of the outcome can then be decomposed as

$$\begin{aligned}\mathrm{Var}(\mathbf{Y}) &= \mathrm{Var}(BX) + \mathrm{Var}(E)\\ &= V_Q + V_R.\end{aligned}$$

PCEV: Statistical model

The PCEV framework seeks a linear combination $w^T\mathbf{Y}$ such that the proportion of variance explained by X is maximised; this quantity is defined as

$$h(w) = \frac{w^T V_Q w}{w^T(V_Q + V_R)w}.$$

PCEV: Statistical model

The PCEV framework seeks a linear combination $w^T\mathbf{Y}$ such that the proportion of variance explained by X is maximised; this quantity is defined as

$$h(w) = \frac{w^T V_Q w}{w^T(V_Q + V_R)w}.$$

Note: When the covariates X are genotypes, $h(w)$ is simply the *heritability* of the linear combination $w^T\mathbf{Y}$.

PCEV: Estimation and Inference

PCEV: Estimation and Inference

- As with PCEV, there is an exact solution, obtained from the *generalised eigenvalue problem*

$$V_Q w = \lambda V_R w.$$

PCEV: Estimation and Inference

- As with PCEV, there is an exact solution, obtained from the *generalised eigenvalue problem*

$$V_Q w = \lambda V_R w.$$

- To perform the association test, we propose two procedures:

PCEV: Estimation and Inference

- As with PCEV, there is an exact solution, obtained from the *generalised eigenvalue problem*

$$V_Q w = \lambda V_R w.$$

- To perform the association test, we propose two procedures:
    - Permutation test

PCEV: Estimation and Inference

- As with PCEV, there is an exact solution, obtained from the *generalised eigenvalue problem*

$$V_Q w = \lambda V_R w.$$

- To perform the association test, we propose two procedures:
  - Permutation test
  - Exact test: Wilks' lambda (one covariate) and Roy's largest root (multiple covariates)

PCEV and PCH

PCEV and PCH

- We did not invent PCEV: it was first introduced by Ott & Rabinowicz (1999) as an alternative to PCA in family-based studies.

PCEV and PCH

- We did not invent PCEV: it was first introduced by Ott & Rabinowicz (1999) as an alternative to PCA in family-based studies.
    - They named this approach *Principal Component of Heritability*. We are advocating for a change of name, since the covariates X do not have to be genetic data.

PCEV and PCH

- We did not invent PCEV: it was first introduced by Ott & Rabinowicz (1999) as an alternative to PCA in family-based studies.
    - They named this approach *Principal Component of Heritability*. We are advocating for a change of name, since the covariates X do not have to be genetic data.
- On the other hand, we are introducing a simple exact test for association tests. Previously, only complicated testing procedures requiring resampling and sample splitting were available.

PCEV: Summary

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.
- **Output**:

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.
- **Output**:
  1. The component maximising the proportion of variance explained by the covariates.

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.
- **Output**:
  1. The component maximising the proportion of variance explained by the covariates.
  2. A set of weights (or loadings), one for each phenotype.

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.
- **Output**:
  1. The component maximising the proportion of variance explained by the covariates.
  2. A set of weights (or loadings), one for each phenotype.
  3. A measure of variable importance: one for each phenotype. This is defined as the **correlation** between a single outcome and the component (in absolute value).

PCEV: Summary

- **Input**: a set of phenotypes values $\mathbf{Y}$ and a set of covariates X.
- **Output**:
  1. The component maximising the proportion of variance explained by the covariates.
  2. A set of weights (or loadings), one for each phenotype.
  3. A measure of variable importance: one for each phenotype. This is defined as the **correlation** between a single outcome and the component (in absolute value).
  4. A p-value for the association between the PCEV and the covariates.

ARCTIC study

ARCTIC study

- Assessment of Risk for Colorectal cancer Tumors In Canada

ARCTIC study

- Assessment of Risk for Colorectal cancer Tumors In Canada
- Data provided by Brent Zanke and Thomas Hudson (Ontario Institute for Cancer Research)

ARCTIC study

- Assessment of Risk for Colorectal cancer Tumors In Canada
- Data provided by Brent Zanke and Thomas Hudson (Ontario Institute for Cancer Research)
- 2,200 individuals (cases and controls) from the Ontario Familial Colon Cancer Registery (OFCCR)

ARCTIC study

- Assessment of Risk for Colorectal cancer Tumors In Canada
- Data provided by Brent Zanke and Thomas Hudson (Ontario Institute for Cancer Research)
- 2,200 individuals (cases and controls) from the Ontario Familial Colon Cancer Registery (OFCCR)
- Methylation levels were measured on lymphocytes derived from whole-blood using the Illumina 450k array.

ARCTIC study

- Assessment of Risk for Colorectal cancer Tumors In Canada
- Data provided by Brent Zanke and Thomas Hudson (Ontario Institute for Cancer Research)
- 2,200 individuals (cases and controls) from the Ontario Familial Colon Cancer Registery (OFCCR)
- Methylation levels were measured on lymphocytes derived from whole-blood using the Illumina 450k array.

Goal: Investigate the association between methylation levels and cigarette smoking, using a gene-based analysis.

ARCTIC study

ARCTIC study

- We compared two approaches

ARCTIC study

- We compared two approaches
  1. A univariate approach: a gene-wide p-value is obtained by taking the minimum of the univariate p-values, and correcting for multiple testing **at the gene level**.

ARCTIC study

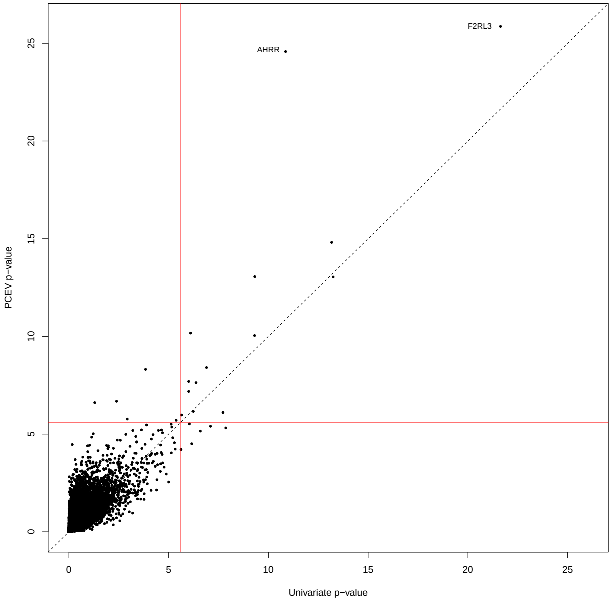
- We compared two approaches
  1. A univariate approach: a gene-wide p-value is obtained by taking the minimum of the univariate p-values, and correcting for multiple testing **at the gene level**.
  2. PCEV: a p-value is obtained using the Wilks' lambda test.

ARCTIC study

- We compared two approaches
  1. A univariate approach: a gene-wide p-value is obtained by taking the minimum of the univariate p-values, and correcting for multiple testing **at the gene level**.
  2. PCEV: a p-value is obtained using the Wilks' lambda test.
- In both cases, we obtain a single p-value per gene.

ARCTIC study

- We compared two approaches
  1. A univariate approach: a gene-wide p-value is obtained by taking the minimum of the univariate p-values, and correcting for multiple testing **at the gene level**.
  2. PCEV: a p-value is obtained using the Wilks' lambda test.
- In both cases, we obtain a single p-value per gene.
- **Note**: we focused on the 1035 control samples, and we considered 18,969 genes, containing between 2 and 607 CpG sites.


F2RL3
AHRR
PCEV p−value
Univariate p−value
0
5
10
15
20
25
0
5
10
15
20
25

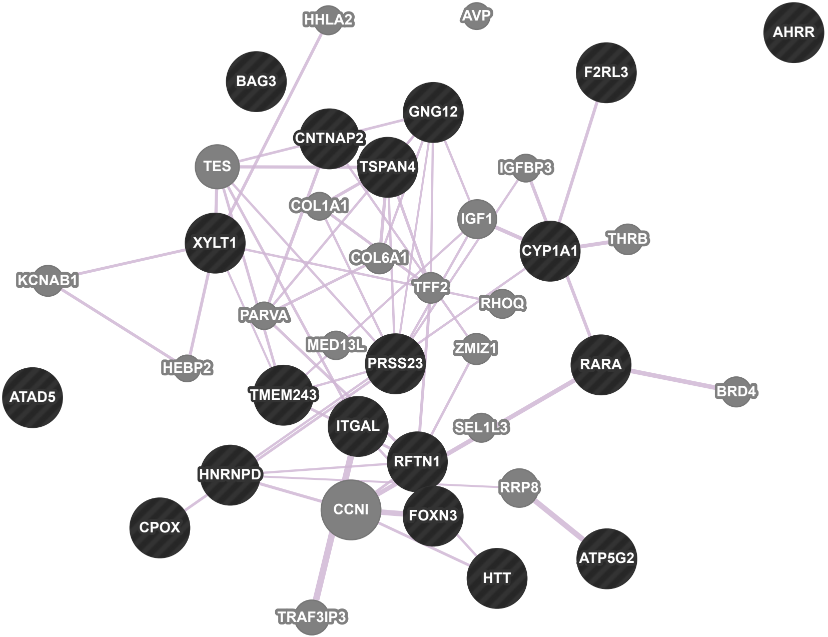
ARCTIC study

- The two most significant genes when using PCEV are F2RL3 and AHRR, respectively. Methylation levels at these two genes are known to be associated with cigarette smoking (Breitling et al., 2011).

ARCTIC study

- The two most significant genes when using PCEV are F2RL3 and AHRR, respectively. Methylation levels at these two genes are known to be associated with cigarette smoking (Breitling et al., 2011).

- The two approaches are generally in agreement, but we see more points above the diagonal than below, suggesting that PCEV has higher power than the univariate approach.


HHLA2
AVP
AHRR
BAG3
F2RL3
GNG12
CNTNAP2
TES
TSPAN4
IGFBP3
COL1A1
IGF1
XYLT1
CYP1A1
THRB
COL6A1
KCNAB1
TFF2
RHOQ
PARVA
MED13L
ZMIZ1
PRSS23
RARA
HEBP2
BRD4
ATAD5
TMEM243
ITGAL
SEL1L3
RFTN1
HNRNPD
RRP8
CCNI
FOXN3
CPOX
HTT
ATP5G2
TRAF3IP3

PCEV: High-dimensional phenotypes

PCEV: High-dimensional phenotypes

- The framework above does not work when $p > n$.

PCEV: High-dimensional phenotypes

- The framework above does not work when $p > n$.

- Regularised versions of the PCEV have been proposed in the literature. However, they all require parameters that are computationally expensive to calibrate.

PCEV: High-dimensional phenotypes

Our main contribution is an extension of PCEV to high-dimensional settings which is

- Simple

PCEV: High-dimensional phenotypes

Our main contribution is an extension of PCEV to high-dimensional settings which is

- Simple
- Computationally very fast

PCEV: High-dimensional phenotypes

Our main contribution is an extension of PCEV to high-dimensional settings which is

- Simple
- Computationally very fast
- Works with $p \gg n$

PCEV: High-dimensional phenotypes

Our main contribution is an extension of PCEV to high-dimensional settings which is

- Simple
- Computationally very fast
- Works with $p \gg n$
- Free of tuning parameters

PCEV: High dimensional outcomes

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the phenotypes (e.g. methylation levels) can be divided in distinct blocks in such a way that:

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the phenotypes (e.g. methylation levels) can be divided in distinct blocks in such a way that:
  - Phenotypes **within** blocks are correlated;

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the phenotypes (e.g. methylation levels) can be divided in distinct blocks in such a way that:
  - Phenotypes **within** blocks are correlated;
  - Phenotypes **between** blocks are uncorrelated.

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the phenotypes (e.g. methylation levels) can be divided in distinct blocks in such a way that:
  - Phenotypes **within** blocks are correlated;
  - Phenotypes **between** blocks are uncorrelated.
- If the size of each block is small enough, we can perform PCEV on each of them, resulting in a single PCEV for each block.

PCEV: High dimensional outcomes

We propose a **block approach** to the computation of PCEV in the presence of high-dimensional outcomes.

- Suppose the phenotypes (e.g. methylation levels) can be divided in distinct blocks in such a way that:
    - Phenotypes **within** blocks are correlated;
    - Phenotypes **between** blocks are uncorrelated.
- If the size of each block is small enough, we can perform PCEV on each of them, resulting in a single PCEV for each block.
- Treating all these "partial" PCEVs as a new, multidimensional pseudo-phenotype, we can perform PCEV again; the result is a linear combination of the original phenotypes **Y**.

PCEV: High dimensional outcomes

Under the above assumption, this is **mathematically** equivalent to performing PCEV in a single-step.

$$\begin{tikzcd}
\{Y_{11},\ldots,Y_{1p_1}\},\ldots,\{Y_{q1},\ldots,Y_{qp_q}\} \arrow[r, leftrightarrow] \arrow[d, "\boxed{\begin{array}{c}\text{Maximize}\\ \text{heritability}\end{array}}"] & X \\
\tilde{Y}_1,\ldots,\tilde{Y}_q \arrow[r, leftrightarrow, thick] \arrow[d, "\boxed{\begin{array}{c}\text{Maximize}\\ \text{heritability}\end{array}}"] & X \\
\tilde{Y} \arrow[r, leftrightarrow, ultra thick] & X
\end{tikzcd}$$

Methyl-seq study

Methyl-seq study

Let's come back to the motivating example.

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)
    - T cells (n=19)

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)
    - T cells (n=19)
    - Monocytes (n=13)

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)
    - T cells (n=19)
    - Monocytes (n=13)
- 5,986 CpG sites

Methyl-seq study

Let's come back to the motivating example.

- BLK gene, located on chromosome 8
- Data provided by Tomi Pastinen (McGill)
- DNA methylation levels derived from bisulfite sequencing
- 40 cell-separated samples, from 3 different cell types
    - B cells (n=8)
    - T cells (n=19)
    - Monocytes (n=13)
- 5,986 CpG sites

Goal: Investigate the association between methylation levels in the BLK region (phenotypes) and cell type (covariate: B cell v. T cell and monocytes)

Results: methylation around BLK

Results: methylation around BLK

- Blocks are defined using physical distance: CpGs within 500kb are grouped together, and then large blocks are split into smaller ones so that no block contains more than 30 CpG sites.

Results: methylation around BLK

- Blocks are defined using physical distance: CpGs within 500kb are grouped together, and then large blocks are split into smaller ones so that no block contains more than 30 CpG sites.
    - 980 blocks were analysed.

Results: methylation around BLK

- Blocks are defined using physical distance: CpGs within 500kb are grouped together, and then large blocks are split into smaller ones so that no block contains more than 30 CpG sites.
  - 980 blocks were analysed.
- Using PCEV, we obtained a single p-value, which is $\approx 4 \times 10^{-4}$ (using 10,000 permutations).

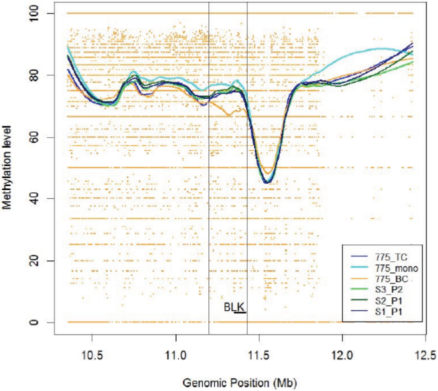
Results: methylation around BLK

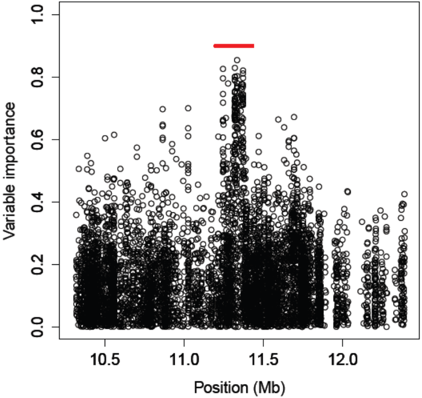
- Blocks are defined using physical distance: CpGs within 500kb are grouped together, and then large blocks are split into smaller ones so that no block contains more than 30 CpG sites.
    - 980 blocks were analysed.
- Using PCEV, we obtained a single p-value, which is $\approx 4 \times 10^{-4}$ (using 10,000 permutations).
- Hence, a single test for all 5,986 variables, and no tuning parameter was required.

Results: methylation around BLK

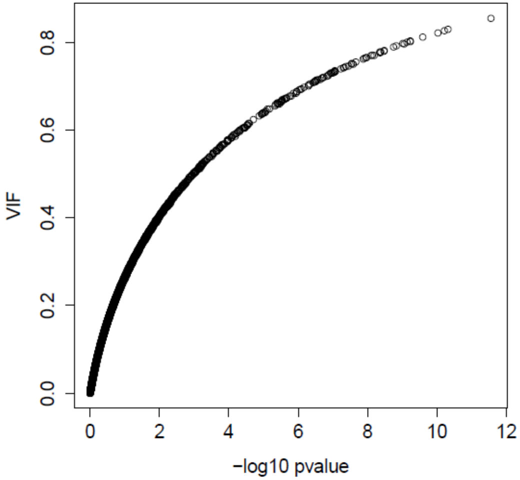
- Blocks are defined using physical distance: CpGs within 500kb are grouped together, and then large blocks are split into smaller ones so that no block contains more than 30 CpG sites.
  - 980 blocks were analysed.
- Using PCEV, we obtained a single p-value, which is $\approx 4 \times 10^{-4}$ (using 10,000 permutations).
- Hence, a single test for all 5,986 variables, and no tuning parameter was required.

We rank the contribution of each CpG sites to this global association using a **Variable Importance Factor**.


Methylation level
0
20
40
60
80
100
10.5
11.0
11.5
12.0
12.5
Genomic Position (Mb)
BLK
775_TC
775_mono
775_BC
S3_P2
S2_P1
S1_P1


Variable importance
0.0
0.2
0.4
0.6
0.8
1.0
10.5
11.0
11.5
12.0
Position (Mb)

Variable importance



Conclusion

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA:

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA:
  - It is **optimal** in capturing the association with covariates.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA:
    - It is **optimal** in capturing the association with covariates.
- Our block approach is a **simple**, computationally **fast** way of handling **high-dimensional phenotypes**.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA:
  - It is **optimal** in capturing the association with covariates.
- Our block approach is a **simple**, computationally **fast** way of handling **high-dimensional phenotypes**.
  - It does not require any tuning parameter.

Conclusion

- Data summary is an important feature in data analysis, and this can be achieved using dimension reduction techniques.
- Principal Component of Explained Variance is an interesting alternative to PCA:
  - It is **optimal** in capturing the association with covariates.
- Our block approach is a **simple**, computationally **fast** way of handling **high-dimensional phenotypes**.
  - It does not require any tuning parameter.
- Simulations and data analyses confirm its advantage over a more traditional approach using PCA.

Software

I remind you that this statistical method is available as an `R` package:

`https://github.com/GreenwoodLab/pcev`

The example on the BLK gene is also included in the vignette accompanying the package.

Questions, comments and/or suggestions?