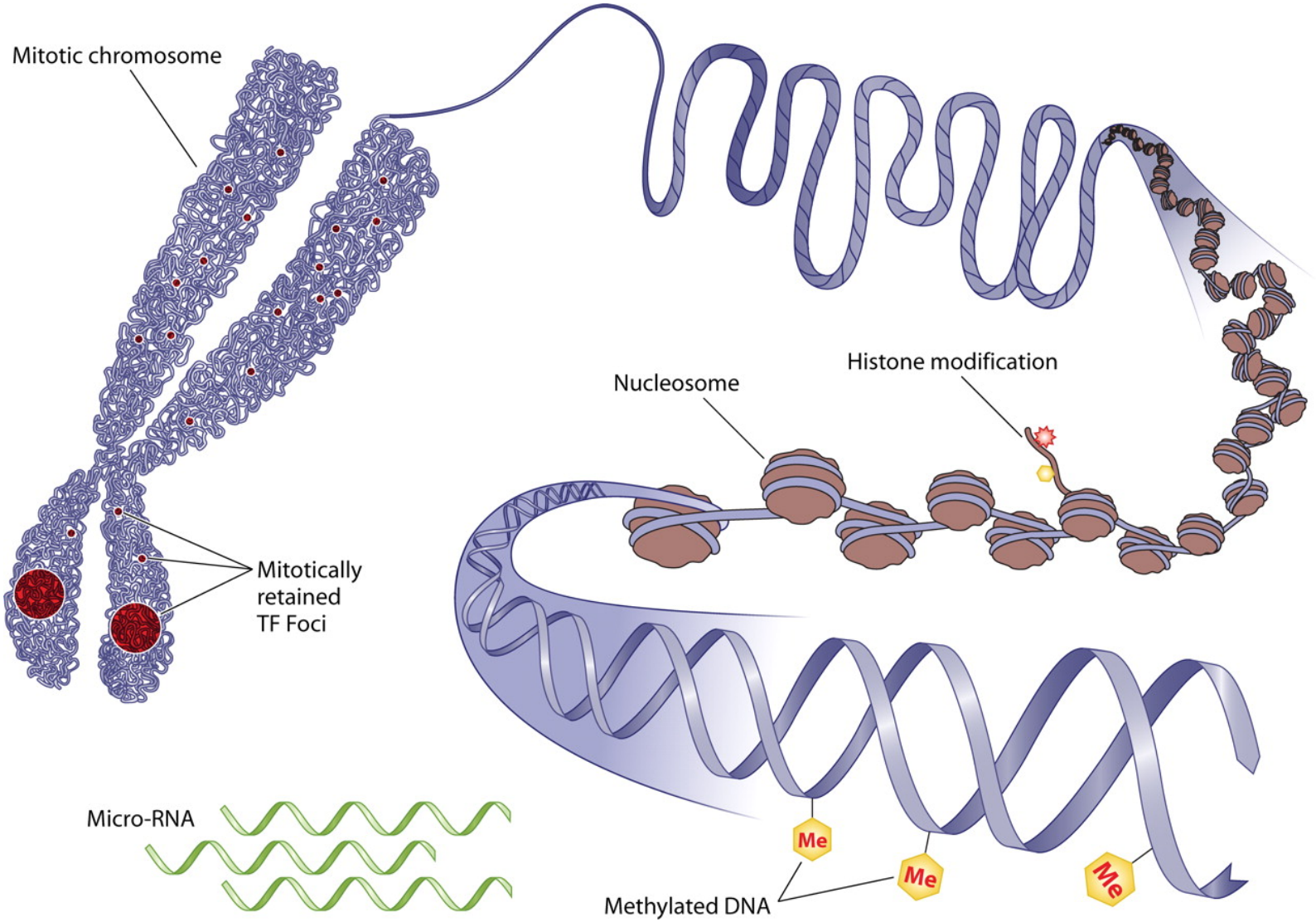
Weierstrass Institute for
Applied Analysis and Stochastics

Simultaneous statistical inference for epigenetic data

Konstantin Schildknecht, Sven Olek, Thorsten Dickhaus

Content

What is Epigenetics?



Source: http://commonfund.nih.gov/epigenomics/figure

What is Epigenetics?

- Central principle of gene regulation.
- Occurs directly on the DNA without changing the sequence.
- DNA Methylation is the selective addition of a methyl group to a cytosine phosphate guanine (CpG) dinucleotide.

What is Epigenetics?

- Central principle of gene regulation.
- Occurs directly on the DNA without changing the sequence.
- DNA Methylation is the selective addition of a methyl group to a cytosine phosphate guanine (CpG) dinucleotide.

And what is it good for?

- Often If a CpG is methylated in proximity of coding regions, Gene silencing occurs.
- Demethylation indicates ability of gene expression.
- Can be used for discrimination of cells, e. g., cancer versus healthy tissue, discriminating immune cells in a blood sample.
- Methylation Signatures, can be used for quantification of cell types.

Observations in methylation studies

For each locus $1 \leq \ell \leq d$ a methylation ratio per observational unit is calculated as

$$X^{(\ell)} = \frac{M^{(\ell)}}{M^{(\ell)} + U^{(\ell)}},$$

where $M^{(\ell)}$ ($U^{(\ell)}$) is an intensity value for the amount of methylated (unmethylated) cells at locus ℓ.

Observations

Assume two groups A and B. Each consist of n_A and n_B, respectively, independent observations that are realizations of iid. d-dimensional random vectors

$$\mathbf{X}_{ik} = (X_{ik}^{(1)}, \ldots, X_{ik}^{(d)})^\top,$$

where $i \in \{A, B\}$ and $1 \leq k \leq n_i$. Assume that the random vectors follow a distribution $\mathcal{L}(X_{A1}) = P$ and $\mathcal{L}(X_{B1}) = Q$. Let $N = n_A + n_B$.

Relative effect

- Let X_A and X_B denote two stochastically independent random variables on a common probability space with probability measure $\mathbb{P}$.
- Assume that X_A and X_B have non-degenerate continuous distributions and denote the normalized cdfs by F_A and F_B.
- The relative effect of F_A with respect to F_B is defined as

$$p_{AB} = \mathbb{P}(X_A < X_B) + \frac{1}{2}\mathbb{P}(X_A = X_B) = \int F_A dF_B.$$

- For a d-variate distribution define the component-wise relative effect for each locus $1 \leq \ell \leq d$ by

$$p_{AB}^{(\ell)} = \int F_A^{(\ell)} dF_B^{(\ell)}.$$

- Let $\mathbf{p}_{AB} = (p_{AB}^{(1)}, \ldots, p_{AB}^{(d)})^\top$ denote the vector of marginal relative effects.

Hypotheses

The functional p_{AB} is capturing central tendencies. We are interested in the hypotheses

$$H_\ell : p_{AB}^{(\ell)} = 1/2 \text{ vs. } K_\ell : p_{AB}^{(\ell)} \neq 1/2,$$

for $1 \leq \ell \leq d$.

System of hypotheses

Let $S \subseteq \{1, \ldots, d\}$. In the remainder, we make use of the notation

$$H_S = \bigcap_{\ell \in S} H_\ell, \quad H_0 = H_{\{1,\ldots,d\}} = \bigcap_{\ell=1}^{d} H_\ell,$$

and refer to H_0 as the global hypothesis in $\mathcal{H}$.

Empirical relative effect

The empirical counterpart of the vector $\mathbf{p}_{AB}$ of relative effects is denoted by $\hat{\mathbf{p}}_{AB} = (\hat{p}_{AB}^{(1)}, \ldots, \hat{p}_{AB}^{(d)})^\top$ with $\hat{p}_{AB}^{(\ell)} = \int \hat{F}_A^{(\ell)} d\hat{F}_B^{(\ell)}$, $1 \leq \ell \leq d$, where $\hat{F}_i^{(\ell)}$ denotes the normalized empirical cumulative distribution function in group $i \in \{A, B\}$ in coordinate ℓ.

Empirical relative effect

The empirical counterpart of the vector $\mathbf{p}_{AB}$ of relative effects is denoted by $\hat{\mathbf{p}}_{AB} = (\hat{p}_{AB}^{(1)}, \ldots, \hat{p}_{AB}^{(d)})^\top$ with $\hat{p}_{AB}^{(\ell)} = \int \hat{F}_A^{(\ell)} d\hat{F}_B^{(\ell)}$, $1 \leq \ell \leq d$, where $\hat{F}_i^{(\ell)}$ denotes the normalized empirical cumulative distribution function in group $i \in \{A, B\}$ in coordinate ℓ.

Theorem 3.5 in Brunner et al. (2002)

Let $V_N \in \mathbb{R}^{d \times d}$ denote the matrix with entries

$$\begin{aligned} v_N^{(\ell,r)} &= \frac{N}{n_A} c_A^{(\ell,r)} + \frac{N}{n_B} c_B^{(\ell,r)}, \\ c_i^{(\ell,r)} &= \text{Cov}\left(Y_{i1}^{(\ell)}, Y_{i1}^{(r)}\right), i \in \{A, B\}, \end{aligned}$$

with the transformed random variables $Y_{Ak}^{(\ell)} = F_B^{(\ell)}(X_{Ak}^{(\ell)})$ and $Y_{Bk}^{(\ell)} = F_A^{(\ell)}(X_{Bk}^{(\ell)})$. Assuming that V_N converges to a positive definite covariance matrix V as $N \to \infty$, it holds that

$$T_d = \sqrt{N}(\hat{\mathbf{p}}_{AB} - \mathbf{p}_{AB}) \xrightarrow{d} \mathcal{N}_d(0, V), \quad N \to \infty.$$

Estimating V_N

Let $\hat{Y}_{Ak}^{(\ell)} = \hat{F}_B^{(\ell)}(X_{Ak}^{(\ell)})$ and $\hat{Y}_{Bk}^{(\ell)} = \hat{F}_A^{(\ell)}(X_{Bk}^{(\ell)})$. For each element (ℓ, r) of V_N let

$$\hat{v}_{N,i}^{(\ell,r)} = \frac{N}{n_i(n_i-1)} \sum_{k=1}^{n_i} (\hat{Y}_{ik}^{(\ell)} - \hat{\bar{Y}}_{i\cdot}^{(\ell)})(\hat{Y}_{ik}^{(r)} - \hat{\bar{Y}}_{i\cdot}^{(r)}),$$

where $\hat{\bar{Y}}_{i\cdot}^{(\ell)}$ denotes the mean of the $\hat{Y}_{ik}^{(\ell)}, k = 1, \ldots, n_i; i \in \{A, B\}$. Note that $\hat{F}_B^{(\ell)}(X_{Ak}^{(\ell)}) = 1/n_B(R_{Ak}^{(\ell)} - R_{Ak}^{(A,\ell)})$, where $R_{Ak}^{(\ell)}$ is the rank of $X_{Ak}^{(\ell)}$ among all N observations in the ℓ-th locus and $R_{Ak}^{(A,\ell)}$ denotes the rank of $X_{Ak}^{(\ell)}$ among all n_A observations.

Wald-type statistic

Making use of a consistent estimator $\hat{V}_N$ defined via the ranks of the observations and a Studentization by $\hat{V}_N$, it follows by Slutsky's lemma that, under $H_0 : \mathbf{p}_{AB} = \mathbf{1}_d/2$, the statistic

$$W_N = N \left(\hat{\mathbf{p}}_{AB} - \frac{1}{2}\mathbf{1}_d \right)^\top \hat{V}_N^{-1} \left(\hat{\mathbf{p}}_{AB} - \frac{1}{2}\mathbf{1}_d \right)$$

is asymptotically χ_d^2-distributed as $N \to \infty$.

- Let π denote an arbitrary but fixed permutation of the set $\{1, \ldots, N\}$.
- Let $\mathbf{X}^\pi = (\mathbf{X}^\pi_{A1}, \ldots, \mathbf{X}^\pi_{An_A}, \mathbf{X}^\pi_{B1}, \ldots, \mathbf{X}^\pi_{Bn_B})$ be the matrix containing the permuted observation vectors from $\mathbf{X} = (\mathbf{X}_{A1}, \ldots, \mathbf{X}_{An_A}, \mathbf{X}_{B1}, \ldots, \mathbf{X}_{Bn_B})$.
- Let the first n_A columns of $\mathbf{X}$ and $\mathbf{X}^\pi$ correspond to group A and the remaining n_B columns to group B.
- Denote by $\tau = \tau(\pi, n_A, n_B)$ the fraction of observations from group B within the first n_A columns of $\mathbf{X}^\pi$.
- Let $\mathbf{p}^\pi_{AB} = \mathbf{p}_{A'B'}$, where $\mathcal{L}(\mathbf{X}_{A'1}) = \tau Q + (1-\tau)P = P'$ and $\mathcal{L}(\mathbf{X}_{B'1}) = (n_A/n_B)\tau P + (1-(n_A/n_B)\tau)Q = Q'$.
- Let $\hat{\mathbf{p}}^\pi_{AB} = \hat{\mathbf{p}}_{A'B'}(\mathbf{X}^\pi)$ denote the estimator of the vector of relative effects based on the permuted data set $\mathbf{X}^\pi$. A simple calculation yields that

$$\mathbf{p}^\pi_{AB} = \tau(1 + n_A/n_B)\mathbf{1}_d/2 + [(1 - \tau(1 + n_A/n_B)]\mathbf{p}_{AB}.$$

- Finally, let

$$\mathfrak{p}^\pi_{AB} = \tau(1 + n_A/n_B)\mathbf{1}_d/2 + [(1 - \tau(1 + n_A/n_B)]\hat{\mathbf{p}}_{AB}.$$

Theorem

Under the general setup from above, assume that the sample sizes n_A and n_B fulfill certain regularity assumptions as $N \to \infty$. Define the statistic

$$W_N^\pi = N(\hat{\mathbf{p}}_{AB}^\pi - \mathfrak{p}_{AB}^\pi)^\top \left(\hat{V}_N^\pi\right)^{-1} (\hat{\mathbf{p}}_{AB}^\pi - \mathfrak{p}_{AB}^\pi), \tag{1}$$

where $\hat{V}_N^\pi$ denotes the estimator from above applied to $\mathbf{X}^\pi$.
Then, the permutation distribution of W_N^π (i. e., its discrete distribution induced by letting π be uniformly distributed on all $N!$ possible permutations of the set $\{1, \ldots, N\}$, while keeping the data $\mathbf{X}$ fixed), the cdf of which we denote by $\hat{R}_N^W$, satisfies

$$\forall t \in [0, \infty) : |\hat{R}_N^W(t) - F\chi_d^2(t)| \xrightarrow{P} 0, \quad N \to \infty.$$

Remark

A result in one dimension was developed in Neubert (2007), in which the statistic was centered by $1/2$.

family wise error rate

For given distributions P and Q, the FWER of φ is defined as the probability under (P, Q) of at least one type I error. The multiple test φ is said to control the FWER strongly at a given level $\alpha \in (0, 1)$, if $\text{FWER}_{(P,Q)}(\varphi) \leq \alpha$ for all possible pairs (P, Q).

family wise error rate

For given distributions P and Q, the FWER of φ is defined as the probability under (P, Q) of at least one type I error. The multiple test φ is said to control the FWER strongly at a given level $\alpha \in (0, 1)$, if $\text{FWER}_{(P,Q)}(\varphi) \leq \alpha$ for all possible pairs (P, Q).

Closed test principle

Add to the set of hypotheses of interest $\{H_\ell | 1 \leq \ell \leq d\}$ all possible intersection hypotheses H_S. Tests every intersection hypothesis H_S at full level α by an arbitrarily chosen level α test φ_S where $S \in 2^{\{1,\ldots,d\}}$.

Only those locus-specific hypotheses H_ℓ are rejected for which all intersection hypotheses H_S with $\ell \in S$ have been rejected by φ_S.

family wise error rate

For given distributions P and Q, the FWER of φ is defined as the probability under (P, Q) of at least one type I error. The multiple test φ is said to control the FWER strongly at a given level $\alpha \in (0, 1)$, if $\text{FWER}_{(P,Q)}(\varphi) \leq \alpha$ for all possible pairs (P, Q).

Closed test principle

Add to the set of hypotheses of interest $\{H_\ell | 1 \leq \ell \leq d\}$ all possible intersection hypotheses H_S. Tests every intersection hypothesis H_S at full level α by an arbitrarily chosen level α test φ_S where $S \in 2^{\{1,\ldots,d\}}$.
Only those locus-specific hypotheses H_ℓ are rejected for which all intersection hypotheses H_S with $\ell \in S$ have been rejected by φ_S.

Remark

We can reduce the d-dimensional vector of test statistics to a subvector which only contains the indices in the subset S to which φ_S refers.
In the corollaries, only the degrees of freedom of the asymptotic χ^2-distributions have to be changed from d to $|S|$.

Model

- *Let the marginals* $\beta(a_i^{(\ell)}, b_i^{(\ell)})$ *distributed.*
- *Set* $b_i^{(\ell)} = 4$, $a_i^{(\ell)} = 3$ *for true nulls, and* $a_B^{(\ell)} = 3 + \delta$ *for alternatives.*
- $\delta = \{0.5, 1, 1.5, 2, 2.5, 3\}$, $d \in \{2, 5, 10\}$.
- *The dependency is modeled by Gaussian correlation of* $AR(1)$ *structure of different degrees.*

Empirical type one error

Tthe empirical type one error was computed as the relative frequency of the occurence of a type I error.

Empirical FWER

Empirical values of the occurence of at least one type one error (FWER) were calculated as, i. e.,

$$\widehat{\text{FWER}} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\{\exists j \leq d_0 : \varphi_{j(k)}(\mathbf{x}^{(\mathbf{k})}) = \mathbf{1}\},$$

where $\varphi_{(k)} = (\varphi_{1(k)}, \ldots, \varphi_{d(k)})^\top$ stands for the multiple test in the k-th simulation run for K repetitions.

Notation

Denote the asymptotic χ^2 multiple test by "χ^2" and the multiple permutation test by "Perm".

Monte Carlo simulation results, based on $K = 10{,}000$ repetitions, regarding the type I error rate for testing the global hypothesis; $(n_A = 20, n_B = 30)$; $\alpha = 5\%$

	$d = 2$		$d = 5$		$d = 10$	
ρ	χ^2	Perm	χ^2	Perm	χ^2	Perm
0	0.0654	0.0428	0.1034	0.0432	0.2154	0.0480
0.2	0.0668	0.0432	0.1092	0.0478	0.2064	0.0408
0.4	0.0730	0.0488	0.1092	0.0482	0.2092	0.0476
0.6	0.0654	0.0426	0.1012	0.0494	0.1898	0.0468
0.8	0.0628	0.0460	0.0848	0.0410	0.1662	0.0448

Monte Carlo simulation results, based on $K = 10{,}000$ repetitions, regarding the type I error rate for testing the global hypothesis; ($n_A = 100, n_B = 150$), $\alpha = 5\%$.

	$d = 2$		$d = 5$		$d = 10$	
ρ	χ^2	Perm	χ^2	Perm	χ^2	Perm
0	0.0527	0.0464	0.0604	0.0448	0.0734	0.0460
0.2	0.0551	0.0456	0.0554	0.0396	0.0772	0.0500
0.4	0.0543	0.0453	0.0590	0.0440	0.0792	0.0476
0.6	0.0520	0.0440	0.0526	0.0396	0.0708	0.0458
0.8	0.0547	0.0486	0.0585	0.0460	0.0640	0.0468

Monte Carlo simulation results, based on $K = 5{,}000$ repetitions, regarding the FWER; $\alpha = 5\%$.

	d_1	0	1	2	3	4
$n_A = 20, \rho = 0.1$	χ^2	0.050	0.056	0.060	0.061	0.065
$n_B = 30, \delta = 3$	Perm	0.021	0.024	0.032	0.036	0.049
$n_A = 20, \rho = 0.5$	χ^2	0.046	0.045	0.045	0.035	0.026
$n_B = 30, \delta = 1$	Perm	0.021	0.016	0.018	0.017	0.011
$n_A = 100, \rho = 0.5$	χ^2	0.028	0.030	0.033	0.029	0.024
$n_B = 150, \delta = 0.5$	Perm	0.020	0.020	0.024	0.022	0.018

Obtained p-values of the tests for relative effects for the loci selected after the screening stage based on the statistic of "χ^2" and "Perm" by applying the closure principle, i. e., for the locus ℓ the smallest significance level such that H_ℓ is rejected.

Locus	$cg00645579$	$cg00974864$	$cg02679745$	$cg08044694$	$cg09134726$
χ^2	0.0046	0.0002	0.00024	0.0002	0.0002
Perm	0.0126	0.0029	0.0029	0.0029	0.0029

Locus	$cg09303642$	$cg09305224$	$cg20070090$	$cg24427660$	$cg24777950$
χ^2	0.0002	0.0047	0.0001	0.0002	0.0002
Perm	0.0029	0.0146	0.0029	0.0076	0.0029

Multiplicity-adjusted p-values of the tests for relative effects with respect to disease groups for three different immune-relevant parameters based on ton the statistic of "χ^2" and "Perm" in combination with the closure principle. The multiplicity-adjusted p-value for parameter ℓ denotes the smallest significance level such that H'_ℓ is rejected for the actually observed data.

Parameter		Treg	tTL	immunoCRIT
Cancer indicator: Healthy colon versus colorectal cancer	χ^2	$< 10^{-16}$	4.926×10^{-13}	$< 10^{-16}$
	Perm	0.0001	0.0001	0.0001
Cancerogenesis: Healthy colon versus early stage cancer	χ^2	5.292×10^{-12}	0.0024	$< 10^{-16}$
	Perm	0.0001	0.0044	0.0001
Cancer progression: Early stage cancer versus late stage cancer	χ^2	0.9043	9.710×10^{-5}	0.0002
	Perm	0.9044	0.0005	0.0011

Discussion

- We proposed the relative effect as a functional of interest for the identification of differentially methylated loci.
- We employed a resampling approach and demonstrated type I error and FWER control in simulations.

Thank you for your attention!


Konstantin Schildknecht, Sven Olek, Thorsten Dickhaus

Simultaneous Statistical Inference for Epigenetic Data

PLOS ONE, Vol. 10, No. 5 (2015), e0125587


Edgar Brunner, Ullrich Munzel, Madan L. Puri

The multivariate nonparametric Behrens–Fisher problem

Journal of Statistical Planning and Inference,108(1):37-53, 2002


Eun Yu Chung, Joseph P. Romano

Multivariate and multiple permutation tests

Technical report, Stanford University,2013-05,2013


Thorsten Dickhaus.

Simultaneous statistical inference. With applications in the life sciences..

Springer Berlin, 2014.