Prediction Models for Cancer Risk and Prognosis using Clinical and DNA Methylation Biomarkers: Considerations in Study Design and Model Development

By

Stefan Hannes Graw
M.S., Technical University of Applied Sciences Wildau, 2014
B.Sc., Technical University of Applied Sciences Wildau, 2011

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Devin C. Koestler

Prabhakar Chalise

Byron J. Gajewski

Jeffrey A. Thompson

Danny R. Welch

Date Defended: 19 April 2019

The dissertation committee for Stefan Hannes Graw certifies that this is the approved version of the following dissertation:

Prediction Models for Cancer Risk and Prognosis using Clinical and DNA Methylation Biomarkers: Considerations in Study Design and Model Development

Chair: Devin C. Koestler

Date Approved: 14 May 2019

Abstract

The ability to accurately predict the prognosis for any given disease is of immense value for clinicians and patients. It can dictate and optimize an individual treatment plan for a patient and ultimately improve their quality of life and reduce the financial burden associated with unnecessary treatment. To allow the accurate prediction of disease prognosis, ongoing development of prediction models is of crucial importance. We introduce a novel curated, ad-hoc, feature selection (CAFS) strategy in the context of the Prostate Cancer DREAM Challenge. We demonstrate enhanced prediction performance of overall survival differences in patients with metastatic castration-resistant prostate cancer by applying CAFS and identify clinically important risk-predictors.

With ongoing advancements in the omics field promising molecular biomarkers are being identified in order to facilitate disease prognosis beyond the capability of clinical information. The identification of such biomarkers depends on the examination of omic marks in adequately powered studies. With the goal to assist researchers in study design and planning of epigenome wide association studies of DNA methylation, we present a user-friendly tool, pwrEWAS, for comprehensive power estimation for epigenome-wide association studies. The R package for pwrEWAS is publicly available at GitHub (https://github.com/stefangraw/pwrEWAS) and the web interface is available at https://biostats-shinyr.kumc.edu/pwrEWAS/.

The enormous volume of omic marks requires stringent evaluation to discover combinations of complementary marks that assemble predictive biomarkers. We therefore present a heuristic feature selection approach that allows one to handle such high-dimensional data. Selection Probability Optimization for Feature Selection (SPOFS) is designed to identify an

optimal subset of omic features from among a vast pool of such features, which collectively improves prediction accuracy and form a biomarker. The integration of such biomarkers can then be utilized in the development and improvement of prediction models.

Acknowledgments

"*Build a world where your children are stronger than you ever were." (Anonymous)*

First and foremost, I want to express my deepest gratitude to my parents Andrea and Torsten Graw. They have created *this* world for me with their unrelenting love and limitless support. They have guided me throughout my life and allowed me to achieve goals I could have never dreamed of. They have given me the chance to grow up in the best environment imaginable and made me into the person I am. They will always be a part of me.

I would like to extend my gratitude to my sister Anne Graw. Growing up, she was my partner in crime, and over the years became my biggest critic and biggest supporter in one person by inspiring me to challenge myself. She has always impressed me with her ambition and creativity.

I am extremely grateful for Erin Willard. Her compassionate love has accompanied me in this journey and encouraged me to reach this point. She provided motivation and caring when things got tough and reassured and entertained me with her unique sense of humor. I cannot thank her enough for being my rock throughout this experience.

I am also very thankful for my family and friends. In particular, I would like to thank Alexander Bratsch and Martin Schulze. Their friendship is of immeasurable importance to me and despite a great distance they remain my closest confidants.

I cannot begin to express my sincerest gratitude to my advisor and chair of my dissertation committee, Dr. Devin Koestler. His mentorship has guided me throughout and beyond my doctoral training. His encouragement and ability to inspire and enthuse was indispensable toward my success. Despite his increasing and instrumental role in the Department

of Biostatistics & Data Science, he always made time to meet with me and discuss our research. I could not have wished for a better mentor.

Lastly, I would like to give a special thanks to my dissertation committee, Dr. Jeffrey Thompson, Dr. Byron Gajewski, Dr. Prabhakar Chalise, and Dr. Danny Welch. I offer my sincere appreciation for their continuous support and guidance and generously offering their time to provide me the opportunity to work with them.

Finally, I would like to express my sincere thanks to our Statistical Omics Working Group for their continuous support and feedback.

Dedication

In memory and honor, I dedicate this dissertation to my beloved grandparents,

Rosemarie Graw and Joachim Busse.

Table of Contents

List of Figures

List of Tables

Introduction

The term “cancer” refers to a collection of diseases which is involved in the growth of abnormal cells with the potential to spread throughout the body (NIH, 2015; WHO, 2018). Cancer is the second most common cause of death in the United States of America (WHO, 2018). Approximately 38.4% of all men and women will be diagnosed with cancer, meaning that more than a third of people will be diagnosed with cancer at some point in their life (NIH, 2018). The national expenditures for cancer care were estimated to be $> \$147$ billion in the United States of America in 2017 (NIH, 2018).

Not only can these costs be drastically reduced by early detection and identifying predictors for the prognosis of cancer, but early detection can also save lives and improve quality of life of cancer patients. Early detection of cancer can lower the risk of dying from cancer, as treatments are likely to be more effective and efficient, ensuring a higher potential of cure (WHO, 2007). The treatment of cancer, generally, has been incredibly challenging, as outcomes vary substantially between individuals. As the result of the personal, societal, and economic burden associated with cancer, it is critically important to develop accurate prediction models and identify predictors for prognosis to facilitate individualized care and personalized cancer medicine.

In chapter one of this dissertation, I describe a curated, ad-hoc, feature selection (CAFS) strategy used for the development of a prediction model during the Prostate Cancer DREAM Challenge, a crowd-based competition. Here, I focus exclusively on sub-challenge 1 of this competition, the prediction of overall survival in patients with metastatic castration-resistant prostate cancer (mCRPC). CAFS attempts to maximize the prediction performance of a given model by iteratively including and excluding features, with the goal of identifying clinically

important risk-predictors. Our ensemble-based Cox regression framework with CAFS resulted in strong overall performance for predicting prostate cancer survival and represents a promising approach for future prediction problems.

One limitation of this project is the consideration of only clinical variables, as no genomic, genetic or epigenetic data were available in this competition. However, it has been demonstrated that molecular biomarkers can improve the prediction of patient prognosis (Bramsen et al., 2017; Choi, Park, Yoon, & Ahn, 2017). Epigenome-wide association studies (EWAS) aim to examine epigenetic marks on a genome-wide level and identify epigenetic biomarkers associated with some exposure(s) or phenotype(s), such as cancer. As for any study, it is crucial to assess sample size in EWAS to identify biomarkers with adequate power. However, direct power assessment of EWAS is challenging due to the complex nature of DNA methylation (DNAm) data (Saadati & Benner, 2014; Teschendorff & Relton, 2018), the most studied epigenetic mark. Due to the lack of tools and methods for power evaluation for EWAS, most EWAS are conducted in the absence of formal power analyses.

In chapter two, I present pwrEWAS, a user-friendly tool for comprehensive power estimation for EWAS. With pwrEWAS, I facilitate power estimation for two-group comparisons of DNAm. Power is calculated using a semi-parametric simulation-based approach in which DNAm data is randomly generated from beta-distributions using parameters from one of several different existing DNAm data sets. I illustrate in a hypothetical EWAS the application of pwrEWAS and demonstrate how pwrEWAS can assist researchers in the design and planning of EWAS.

Finally, in chapter three I discuss the integration of epigenetic biomarkers in the development of prediction models. To achieve this goal, I present a heuristic feature selection

approach called Selection Probability Optimization for Feature Selection (SPOFS). SPOFS is designed to identify an optimal subset of omic features from among a vast pool of such features, which collectively improve the ability to predict some outcome or response. Therefore, in an initial step, SPOFS involves filtering data to a manageable number of features. Then, sets of features are iteratively selected and their performance is assessed. The selection probability of each evaluated feature is then optimized based on its individual performance and a new set of features is selected. The fundamental idea is to optimize the selection probability of features to increase the probability of identifying an optimal set. Even though SPOFS was developed for a time-to-event analysis in the context of a nested case-control study, this methodology is applicable to any project that desires a selection of features evaluated on some prediction metric. The prediction performance of SPOFS was benchmarked against the performance of Lasso (Least Absolute Shrinkage and Selection Operator) in a nested case-control study of lung cancer risk, embedded in the Carotene and Retinol Efficacy Trial (CARET), and in a variety of simulated data sets. I demonstrate that SPOFS is a competitive feature selection method with the potential to outperform Lasso in certain scenarios.

1 Chapter 1: An ensemble-based Cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mCRPC) patients

This chapter has previously been published in whole with minor adaptations since publication and is available as an open access article. Meier R and Graw S, et al. An ensemble-based Cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mCRPC) patients. F1000Research 2016, 5:2677 (Meier et al., 2016).

1.1 Abstract

From March through August 2015, nearly 60 teams from around the world participated in the Prostate Cancer Dream Challenge (PCDC). Participating teams were faced with the task of developing prediction models for patient survival and treatment discontinuation using baseline clinical variables collected on metastatic castrate-resistant prostate cancer (mCRPC) patients in the comparator arm of four phase III clinical trials. In total, over 2,000 mCRPC patients treated with first-line docetaxel comprised the training and testing data sets used in this challenge. In this paper we describe: (a) the sub-challenges comprising the PCDC, (b) the statistical metrics used to benchmark prediction performance, (c) our analytical approach, and finally (d) our team's overall performance in this challenge. Specifically, we discuss our curated, ad-hoc, feature selection (CAFS) strategy for identifying clinically important risk-predictors, the ensemble-based Cox proportional hazards regression framework used in our final submission, and the adaptation of our modeling framework based on the results from the intermittent leaderboard rounds. Strong predictors of patient survival were successfully identified utilizing our model building approach. Several of the identified predictors were new features created by our team via strategically

merging collections of weak predictors. In each of the three intermittent leaderboard rounds, our prediction models scored among the top four models across all participating teams and our final submission ranked 9th place overall with an integrated area under the curve (iAUC, see **Appendix I - Integrated area under the curve (iAUC))** of 0.7711 computed in an independent test set. While the prediction performance of teams placing between 2nd-10th (iAUC: 0.7710-0.7789) was better than the current gold-standard prediction model for prostate cancer survival, the top-performing team, FIMM-UTU significantly outperformed all other contestants with an iAUC of 0.7915. In summary, our ensemble-based Cox regression framework with CAFS resulted in strong overall performance for predicting prostate cancer survival and represents a promising approach for future prediction problems.

1.2 Introduction

Today, prostate cancer is one of the most prevalent cancers afflicting men in the Western world. In addition to the prevalence of this disease, the mortality rates for prostate cancer ranked fifth among the most common causes of cancer death worldwide in 2012 (http://www.cancerresearchuk.org/). In the US alone, approximately 137.9 out of 100,000 men were diagnosed with prostate cancer each year from 2008–2012, with an average annual mortality rate of 21.4 out of 100,000 men. (http://www.seer.cancer.gov/statfacts/html/prost.html). According to the Cancer Prevalence and Cost of Care Projections, the total annual cost of prostate cancer in 2016 has been estimated at 14.3 billion dollars (http://www.costprojections.cancer.gov/).

Over the course of the last decade in the US, approximately 15% of prostate cancer cases were initially diagnosed with metastatic disease (stage IV). Androgen deprivation therapy (ADT) is the established treatment for these cases, but one third of patients develop resistance and their

disease progresses to metastatic castrate-resistant prostate cancer (mCRPC) (https://www.synapse.org/ProstateCancerChallenge). Treatment of mCRPC has been historically challenging, and while docetaxel – the current front-line therapy for mCRPC – has been effective at improving mCRPC survival at the population level, a significant fraction of patients do not respond to treatment or prematurely discontinue treatment due to adverse events (AE) (Schallier, Decoster, Braeckman, Fontaine, & Degreve, 2012), leading to substantial variation in the individual outcomes between mCRPC patients. For this reason, and because of the tremendous personal, societal, and economic burden associated with this disease, there is significant interest both in the identification of individual predictors for mCRPC prognosis as well as the development of prognostic models that can be used to identify high-risk mCRPC patients.

In a recent publication (Halabi et al., 2014), Halabi et al. utilized data from a phase III trial consisting of over one thousand mCRPC patients to develop and test a prognostic model for overall survival among patients receiving first-line chemotherapy. The time dependent area under the curve (tAUC) was > 0.73 in both testing and independent validation data sets, suggesting strong performance of the Halabi et al. model for identifying low- and high-risk mCRPC patients. Notwithstanding the significant advances made by Halabi et al. and others toward the development of accurate prognostic models for mCRPC outcomes (Chang et al., 2015; Halabi et al., 2014; van Soest et al., 2015), there remains ample room for improved prediction performance.

Motivated by the potential to further improve existing risk-prediction tools along with growing worldwide burden of prostate cancer, the Prostate Cancer Dream Challenge was launched in March 2015 and included the participation of nearly 60 teams from around the world. The Prostate Cancer Dream Challenge was composed of two distinct sub challenges; in

sub challenge 1, teams competed in the development of prognostic models for predicting overall survival based on baseline clinical variables, whereas the objective of sub challenge 2 involved the development of models to predict short-term treatment discontinuation of docetaxel (< 3 months) due to adverse events (AE). To assist in the development and testing of prediction models, approximately 150 variables collected on over 2,000 mCRPC patients treated with first-line docetaxel in one of four different phase III clinical trials were used. Three of the four trials were combined to generate the training data set, which was used for model-building and development, while data from the remaining trial were withheld from challenge participants and used as an independent test set to evaluate each of the submitted models (Guinney et al., 2017).

In the present manuscript, we focus exclusively on our methodological approach to sub challenge 1. Broadly speaking, the first step of our team's approach to sub challenge 1 involved an initial screening of the data: data cleaning and processing, creation of new variables from existing data, imputation and/or exclusion of variables with missing values, and normalization to standardize the data across trials. The final "cleaned and standardized" training data was then used to fit to an ensemble of multiple Cox proportional hazards regression models whose constituent models were developed using curated, ad-hoc, feature selection (CAFS). Models developed by our team were subjected to internal cross-validation within the training data set to identify instances of model overfitting and to assist in further refinements to our prediction models. The source code utilized for our approach can be accessed via the Team Jayhawks Prostate Cancer Dream Challenge project web page (https://www.synapse.org/#!Synapse:syn4214500/wiki/231706) or directly via the GitHub repository webpage (https://github.com/richard-meier/JayhawksProstateDream).

1.3 Materials and methods

1.3.1 Data

A detailed description of the data sets used in this challenge can be found on the Prostate Cancer Dream Challenge web page (https://www.synpase.org/ProstateCancerChallenge). Briefly, the training set originated from the ASCENT-2 (Novacea, provided by Memorial Sloan Kettering Cancer Center), MAINSAIL (Celgene) and VENICE (Sanofi) trials (Petrylak et al., 2015; Scher et al., 2011; Tannock et al., 2013). For the 1600 patients in the training data, baseline covariate information and clinical outcomes (i.e., time to death and time to treatment discontinuation) were provided to participating teams for the purposes of model development and training. Although baseline covariate information for a subset of patients in the ENTHUSE-33 (AstraZeneca) trial (Fizazi et al., 2013; Tannock et al., 2013) scoring set was provided to participating teams (n = 157), the clinical outcomes for each of these patients were censored and withheld from teams throughout the duration of the challenge. Specifically, the ENTHUSE-33 data set (n = 470) was split into two disjoint sets that consisted of 157 and 313 patients. Whereas an undisclosed randomly selected subset of the 157 patients was used for model evaluation in each intermittent leaderboard round, the remaining 313 patients were withheld completely from participating teams and used only in the final scoring round.

1.3.2 Preprocessing

All aspects of our approach, from data preprocessing to model development and cross-validation, were implemented using R version 3.2.1 (2015-06-18) (https://www.r-project.org/). Baseline covariate information on subjects comprising the training data were reformatted and normalized according to the type of variable (i.e., categorical, ordinal, numeric) and feature type (i.e., medical history, laboratory values, etc.). Cleaned and normalized baseline features were

then used to derive additional novel risk predictors. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/dataCleaningMain.R)

Several groups of binary variables representing patient specific clinical information and prior medical history reported on patients were merged into new features. Three different merging types were explored: “logical or”, regular summation, and z-score weighted summation. For the latter, each individual feature in the training set was fit against survival time with a Cox proportional hazards model and their resulting z-scores were used to derive weights that were proportional to each variable’s strength of association with survival (https://github.com/richard-meier/JayhawksProstateDream/blob/master/deriveHardcodedWeights.R). Summation variables were created for 3 main categories: medical history information, prior medication information and metastasis information. For each of these categories, new variables generated by merging specific subcategories (i.e., protective, harmful, total, visceral, etc.) were created.

A participant’s target lesion volume (TLV) was generated by multiplying each target lesion by its size, followed by summing over all lesions within that participant (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/lesion_volume.R). To impute the TLV for the ASCENT-2 trial, we calculated the average TLV per lesion within individuals who survived or died in the MAINSAIL or VENICE trials, and multiplied these separate averages by the number of non-bone lesions found in the ASCENT-2 data. To classify whether for each category a feature was in the subcategory “protective” or “harmful”, their z-scores, when individually fitting against the outcome, were used. A feature was labeled "protective" if its z-score was greater than 1.64 and "harmful" if its z-score was smaller than -1.64.

Principal component analysis (PCA) was used to split numerical laboratory values into components that best explained their variation (see above: “deriveHardcodedWeights.R”). The top PCs were then treated as new features. In order to address issues or findings involving some specific variables, additional features were created: The Eastern Cooperative Oncology Group (ECOG) performance status score was both included as continuous and categorical variable. Age groups were also recoded as an ordinal age risk variable for which 0 represented patients older than 75 years, 1 represented patients younger than 65 years and 2 represented patients with ages between 65–75 years. The latter was motivated by our observation of a non-linear trend between age and survival time.

Race was recoded into a binary variable where 1 referred to patients labeled as “white” or “other” and 0 represented patients that did not fall into one of those two categories (e.g. "black", "asian", etc.). The features “harm_pro” and “harm_pro2” were created by fitting the summation variables of the medical history subgroups “harmful” and “protective” against the outcome and obtaining the z-scores of these subgroup summation variables. The difference between the two features was that harm_pro exclusively fitted the two summation variables, whereas harm_pro2 also utilized a set of important predictor variables for the initial fit. The two z-score weighted sums (corresponding to the two sets of features utilized for the previously mentioned fit) of these summation variables then correspond to the two new features. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/add_additional_features.R)

1.3.3 Model building and feature selection

Our methodological framework utilized an ensemble of Cox proportional hazards regression models that were found to be individually competitive in predicting survival. For each patient, the ensemble-based risk scores were generated as a weighted sum of the individually

estimated risk scores from separate Cox-regression models, fit using the “coxph” function in the “survival” R-package (Therneau, 2015) (Figure 1C). Feature selection among the competitive risk-prediction models that constituted our ensemble was undertaken by a method we call curated, ad-hoc, feature selection (CAFS). This method attempts to maximize the prediction performance of a given model by iteratively including and excluding features from a baseline initial model. The method is greedy in the sense that in each step of the algorithm, only the model candidates that achieve the current "local best" performance are selected. Each iteration started with a group of experts making two executive decisions based on a set of possible model candidates for which performance was evaluated in prior iterations. First, one model was nominated as the best current model and a decision was made whether to expand or shrink the model, or terminate the procedure and keep the model, in its current form (Figure 1A). Choosing the current best model was guided by a candidate’s estimated performance, performance of the previous best model, as well as knowledge of the researchers as to whether the form and components of a given model were reasonable in the context of the problem at hand. An example for the latter case would be that a newly introduced interaction term between completely unrelated features might be rejected after evaluation, even though it technically achieved the current best performance.

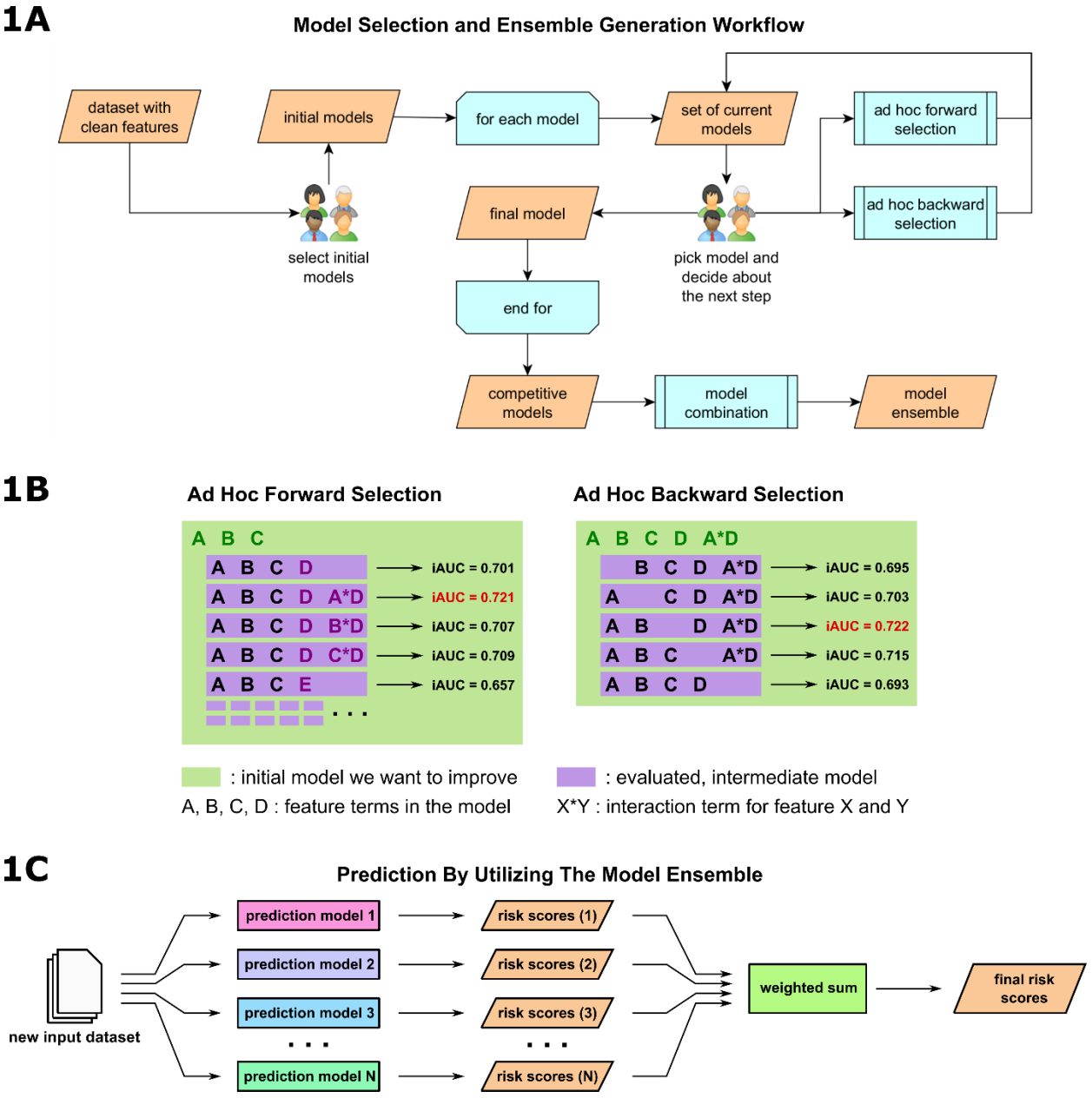


Figure 1 Model building and model ensemble utilization.

(1A) Competitive prediction models were built individually by a curated, ad-hoc feature selection procedure. In each step researchers picked a new best model from the set of current models based on an optimization criterion and decided how it would be processed. (1B) Models were optimized by either forward selection, in which a new feature was added, or backward selection, in which a feature that had become obsolete was removed. Both selection methods generated a set of new models for which performance was predicted via in-depth cross-validation. (1C)

Once a variety of competitive prediction models had been created, models were combined into an ensemble, which averaged their individual predictions in order to increase performance.

Model reduction was done via ad-hoc backward selection (Figure 1B). In this procedure a set of new models was generated by individually excluding each parameter or feature present in the current model. For each of these models, performance was evaluated based on a previously chosen optimization criterion, i.e., integrated time-dependent area under the curve (iAUC). The criterion was estimated via a cross-validation procedure in which the training set was repeatedly split into two random subsets of a fixed size. The first subset was used to estimate parameters of a given model, whereas the second subset was used to predict the outcome using the previously estimated parameters and to calculate the optimization criterion based on comparing the prediction with the true outcome. In our study, we utilized two-thirds for the parameter estimation subset, i.e., first subset, while the remaining one third comprised the second subset. The average of the calculated optimization criterion values, obtained from all random splits, served then as a performance estimate. We used 10,000 cross-validation steps for each model in our study to ensure stability of the average performance. The new models and performance estimates were then used as the basis for subsequent iterations.

Expansion of a model was accomplished using an ad-hoc forward selection procedure (Figure 1B). In this procedure several new models were created for each feature within the feature space. Each subset of new models contained one base model that included only main effect terms for new features, i.e., no interaction terms included. All other models in the subset further expanded this base model by individually introducing an interaction term with each element already in the previous best model.

Performance of each new model was again assessed via the cross-validation procedure. Since this step iterated over the feature space, it created a large amount of different models. To make this step computationally feasible, the number of cross-validation iterations had to be reduced. In our study, 500 cross-validation steps per new model were utilized. (https://github.com/richard-meier/JayhawksProstateDream/blob/master/src/modelTuning.R)

Finally, since the variances of these performance estimates were much higher than in the shrinkage step, the top 30 performing models were chosen, and performance was re-estimated via 10,000-fold cross-validation. This set of new models and performance estimates was then used in the next iteration. Once iterations provided only marginal performance increases, the procedure was terminated, and a final model was declared. Different models for the ensemble were found either by choosing different intermediate models as the current best and branching off a certain path, or by choosing different initial models.

1.3.4 Model evaluation

Each of the sub challenges in the Prostate Cancer Dream Challenge had its own prediction scoring metrics. In sub challenge 1A, participants were asked to submit a global risk score and time dependent risk scores, optimized for 12, 18 and 24 months. These risk scores were evaluated utilizing two scoring metrics: a concordance index (c-index, see **Section 3.3.4**), and an integrated time dependent area under the curve (iAUC; 6–30 months). The time specific risk scores were assessed using AUC's computed using Hung and Chiang's estimator of cumulative AUC (Hung & Chiang, 2010). In sub challenge 1B, participants were asked to predict the time to event (death). The predictions of time to event were scored utilizing the root mean squared error (RMSE), using patients with known days to death.

When applying CAFS, we utilized the iAUC calculated from the predicted risk scores as an optimization criterion. This measure was also used by the challenge organizers for performance assessment in the scoring rounds for sub challenge 1A. While participants were asked to predict the risk score for overall survival based on patients' clinical variables, they were also tasked to predict the time to event (TTE) in sub challenge 1B. We used the risk score for each patient to model the TTE:

$$TTE_i = f(riskScore_i) + \epsilon_i$$

Where $riskScore_i$ corresponds to the risk score calculated in sub challenge 1A for the i^{th} patient and *f* is an unknown smoothing function. We estimated *f* using a Generalized Additive Model (GAM) via the "gam" function within the "mgcv" package in R (Wood, 2011). When regressing TTEs on risk we used only the subset of individuals who died.

1.4 Results

The principal component analysis with all laboratory values revealed that the first principal component (PC) was highly correlated with patient survival. Furthermore, across all laboratory values, only a subset of six features (baseline levels of: albumin, alkaline phosphatase, aspartate aminotransferase, hemoglobin, lactate dehydrogenase and prostate specific antigen) contributed significantly to explaining the variation in said first component. Thus, in the first PC only these six laboratory values were used during model building and development. In addition to the first principal component, several other newly created metavariables were identified as clinically relevant predictors by our model building procedure. Three z-score weighted sums merging metastases locations, medical history and prior medication were included in our prediction models. The "logical or" merged variable, whether or not a patient had any known medical history issues, was also utilized. The protective versus harmful subcategorization was

only included in the models in the form of the sum of protective medical history features. However, this category only included a single feature, vascular disorders (yes/no).

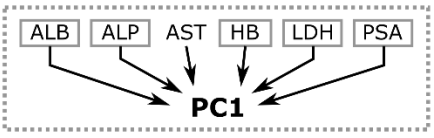
We developed 5 competitive prediction models (M1 – M5) that were used in our Cox proportional hazards regression ensemble (Figure 2). All models were developed by either refining a previous model via CAFS or by building a model from the bottom up via CAFS. M1 used the best model found by manually selecting promising features as its initial model. M2 used an intermediate model from the CAFS procedure of M1 to deliberately branch off and provide a similar, yet different model. M3 and M5 were both built by using an initial model solely utilizing the strong predictors target lesion volume and principal component 1 but branching off in early iterations. M4 was built by using an initial model utilizing target lesion volume and the alkaline phosphatase level under the restriction that principal component 1 was excluded from the feature space.

2A

M1	M2	M3	M4	M5	Feature type	Description of the information type
×	×				AGE	Patient age group (3 groups): years coded as 18-64, 65-74, 75+
		×	×	×	ALP	Alkaline Phosphatase level
			×		AST	Aspartate Aminotransferase level
	×				BISPH	Prior Bisphosphonate medication (Yes / No)
×	×	×	×		ECOG	ECOG patient performance status (0,1,2,3,4)
×	×	×			ESTRO	Prior Estrogen medication (Yes / No)
		×			GLCC	Prior Glucocortocoid medication (Yes / No)
		×	×		HAPRO	Z-score metavariable merging "harmful" and "protective" medical history sums
×			×	×	HB	Hemoglobin level
			×		LDH	Lactate Dehydrogenase level
	×	×	×	×	MHIST_OR	Logical or merged medical history data (asks if any medical history issues are known)
	×	×	×	×	MHIST_ZW	Z-score metavariable merging medical history data
×	×		×		METAS_ZW	Z-score metavariable merging metastases location data
×	×				MI	Myocardial infarction diagnosed in medical history (Yes / No)
×					NA.	Sodium level
	×		×	×	NEU	Neutrophil level
×	×	×		×	PC1	Principal component 1 of the most relevant laboratory values
		×	×	×	PHOS	Phosphorus level
				×	PRIMED_ZW	Z-score metavariable merging prior medication information
		×			PROST	Prostate lesions present (Yes / No)
×					PROT_MH	"Protective" medical history diagnosis (Vascular Disorders: Yes / No)
×	×				RACE	Race information: white (Yes / No)
	×	×	×	×	TLV	Target lesion volume: metavariable merging target lesions by volume
Features only in models as part of PC1					ALB	Albumin level
					PSA	Prostate Specific Antigen level

Feature in 4/5 3/5 2/5 1/5 of models

Feature was also used in the model proposed by Halabi et al. [2]



2B

Model	Parameters	Interaction terms		iAUC (cv)
M1	14	4	AGE*MI, METAS_ZW*RACE, METAS_ZW*NA., METAS_ZW*HB	0.743
M2	15	3	AGE*MI, METAS_ZW*RACE, NEU*BISPH	0.745
M3	12	2	MHIST_OR*PHOS, ALP*GLUCOCORTICOID	0.743
M4	14	2	MHIST_OR*PHOS, ALP*ECOG	0.740
M5	12	4	MHIST_OR*PHOS, TLV*NEU, PC1*ALP, PRIMED_ZW*HB	0.741
Ensemble	38	11		0.757

Figure 2 Generated models utilized in the final challenge submission.

(2A) The ensemble consisted of five different models, M1 to M5, which ended up sharing many feature types even though they were individually generated under different conditions. (2B) All models made use of a similar number of parameters and achieved comparable performance in cross-validation. Performance further increased when using the model ensemble.

While no single feature was utilized in every model M1–M5, five different features were shared between four models, six features between three models, four features between two models and eight features were unique to a model (Figure 2A). Each model had at least one unique feature. Between two and four interaction terms (two-way interaction terms) were present in all the observed models (Figure 2B). One interaction was shared between the models M3, M4 and M5, while two interactions were shared between two models M1 and M2. Including components of newly derived features, eight features that were included in the original model by Halabi et al. in some form, were also utilized in the model ensemble. In total, the ensemble contained 38 coefficients, out of which 11 were pairwise interaction terms across all models.

The estimated iAUC during performance assessment was found to be stable up to approximately three decimals when using 10,000-fold cross-validation. Similar estimated performance within the range of 0.005 iAUC difference was achieved between the competitive prediction models, the highest total iAUC being 0.745. Optimal weights were chosen based on randomly initializing weights 100 times and estimating performance. Performance tended to be optimized the smaller the maximum pairwise difference between weights in an ensemble was. The best possible performance was estimated when choosing equal weights for all models. This ensemble was chosen as the best model. Utilizing the ensemble led to an estimated performance increase of 0.012 iAUC.

During the three leaderboard rounds the team explored and submitted various methodologies. Top performing submissions were always Cox proportional hazards models that outperformed more sophisticated approaches such as generalized boosted regression models and random survival forests. From scoring round 2 onward, single models utilizing CAFS were also submitted. In all intermittent leaderboard rounds, at least one of our submitted entries ranked among the top 4 performing models of sub challenge 1A (Figure 3A). In sub challenge 1B, at least one submission was within the top 3 performing models, with the exception of the second leaderboard round were our best model ranked number 12. Our models achieved performances ranging from 0.792 to 0.808 iAUC in 1A and from 172.51 to 196.25 RMSE in 1B. In the final scoring round, team FIMM-UTU (Guinney et al., 2017) significantly outperformed all other contestants with an iAUC of 0.7915 (Figure 3B). Our submission for 1A that utilized the CAFS ensemble achieved rank 9 with an iAUC of 0.7711. The performances of teams ranking from 2nd to 10th were very similar. While the difference in performance between rank 1 and 2 was 0.0126 iAUC, the difference in performance between our method and rank 2 was only 0.0078 iAUC. Our submitted model ensemble also successfully outperformed the previous best model by Halabi et al. (Halabi et al., 2014), which was placed at rank 36 with an iAUC of 0.7429. Sub challenge 1B was won simultaneously by 6 teams out of which our method achieved rank 3.

3A

Round	Subchallenge	Performance	Achieved Rank
1	1A	iAUC = 0.8081	1
1	1B	RMSE = 183.74	2
2	1A	iAUC = 0.8028	4
2	1B	RMSE = 196.25	12
3	1A	iAUC = 0.7922	3
3	1B	RMSE = 172.51	1

3B

Team	Subchallenge	Performance	Rank	Winners
FIMM-UTU	1A	iAUC = 0.7915	1	Rank 1 only
JayHawks	1A	iAUC = 0.7711	9	—
CAMP	1B	RMSE = 194.41	1	Rank 1 to 6
JayHawks	1B	RMSE = 195.97	3	

Figure 3 Team performance during the challenge.

(3A) Submitted models were consistently ranked at the top of the leaderboards during the scoring rounds before the final submission. Models build via the CAFS procedure were submitted starting with the second leaderboard round. (3B) The final challenge submission made use of the described model ensemble approach and was placed at rank 9 in sub challenge 1A and at rank 3 in sub challenge 1B.

1.5 Discussion

Many feature types present in the original model by Halabi et al. (Halabi et al., 2014) were also independently picked up and retained by CAFS. This solidifies the idea that these might be key components influencing survival. Considering that five out of these eight were also involved in the first principal component, which was one of the strongest predictors, does also support this. Another set of potentially interesting predictors are those shared between three or more models.

It is debatable whether the fact that a lot of overlap exists between the various sub-models points towards the validity of selected features and the developed approach, or a potential bias in the feature selection procedure. However, the former appears more likely in the light of the approach's good performance on new data in the competition.

The included interaction terms are difficult to interpret. There is no guarantee that an interaction is modeling a direct relationship and some terms might be artifacts of higher order interactions or confounding issues. Also, when solely including terms into the model based on the optimization criterion in each step of CAFS, there is a bias to include interaction terms. Since they introduce more parameters into the model than a main effect, they have more opportunity to improve the model within each step, even though including two different main effects in a row might be more beneficial. While our team was aware of this issue and cautious with the selection of sub-models, this still leaves potential for making suboptimal choices. This weakness could potentially be addressed in the future by switching to a parameter count based iteration, rather than a feature type-based iteration.

The performed recoding of the age groups is still problematic. Intuitively, it does not make sense that the order "oldest, youngest, in-between" would be related to the outcome when

disease progression usually worsens with age. A possible explanation might be that the oldest patient group is confounded with a subset of people that are resistant to the disease and have already survived for a long time. Further research is required to validate this effect.

Overall the presented method successfully built a robust predictor for the target outcome. Evidence for this is provided by the fact that the estimated performance via in-depth cross validation (iAUC = 0.757) was close to the reported performance on the larger, final leaderboard set (iAUC = 0.771) and the fact that our models were among the top performing candidates throughout the entire challenge. It should also be highlighted that the required human intervention in each selection step gives the team of researchers a lot of control, which can be very useful to introduce knowledge about the feature space into the selection process, but also limits the reproducibility. An example of this benefit is that despite the pointed-out weakness in the implementation, the team was able to account for it by rejecting inclusions of interactions that did not have a great enough impact. If desirable, early branches of the selection process can be tailored towards features with a known connection to the outcome when multiple feature inclusions provide similar performance benefits.

1.6 Conclusion

The presented method generated a model ensemble that was able to outperform the previous best efforts to predict survival in prostate cancer patients. The developed model ensemble also successfully competed with the top performing research teams in the Prostate Cancer Dream Challenge and was among the winning teams in sub challenge 1B. We attribute this success to careful data cleaning, our efforts to derive novel features and the fact that skeptic, human decision making is integral to each iteration of the curated ad-hoc feature selection. Due to its general applicability to model building, especially in exploratory settings, the approach is

promising in being useful for researchers around the world. Future studies will need to validate the presented, potentially disease associated features and potential weaknesses in the CAFS procedure should be investigated and addressed.

1.7 Data availability

The Challenge data sets can be accessed at:

https://www.projectdatasphere.org/projectdatasphere/html/pcdc

Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, and the clinical trials data dictionary can be found at:

https://www.synapse.org/ProstateCancerChallenge

The code and documentation underlying the method presented in this paper can be found at:

http://dx.doi.org/10.5281/zenodo.49063 (Chalise et al., 2016)

1.8 Contribution

In this project I participated in the conception of the idea of CAFS, the model building process, model assessment and interpretation, the organization and presentation of the code, implementation of data cleaning functions for medication history, combining treatment history variables into new variables, and writing the manuscript.

2 Chapter 2: pwrEWAS: A user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS)

This chapter has previously been published in whole with minor adaptations since publication and is available as an open access article. Stefan Graw, M.Sc.; Rosalyn Henn, B.Sc.; Jeffrey A. Thompson, Ph.D.; Devin C. Koestler, Ph.D. pwrEWAS: A user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS). BMC Bioinformatics 2019. (Graw, Henn, Thompson, & Koestler, 2019)

2.1 Abstract

When designing an epigenome-wide association study (EWAS) to investigate the relationship between DNA methylation (DNAm) and some exposure(s) or phenotype(s), it is critically important to assess the sample size needed to detect a hypothesized difference with adequate statistical power. However, the complex and nuanced nature of DNAm data makes direct assessment of statistical power challenging. To circumvent these challenges and to address the outstanding need for a user-friendly interface for EWAS power evaluation, we have developed pwrEWAS. The current implementation of pwrEWAS accommodates power estimation for two-group comparisons of DNAm (e.g. case vs control, exposed vs non-exposed, etc.), where methylation assessment is carried out using the Illumina Human Methylation BeadChip technology. Power is calculated using a semi-parametric simulation-based approach in which DNAm data is randomly generated from beta-distributions using CpG-specific means and variances estimated from one of several different existing DNAm data sets, chosen to cover the most common tissue-types used in EWAS. In addition to specifying the tissue type to be used for DNAm profiling, users are required to specify the sample size, number of differentially

methylated CpGs, effect size(s) (Δ_β), target false discovery rate (FDR) and the number of simulated data sets, and have the option of selecting from several different statistical methods to perform differential methylation analyses. pwrEWAS reports the marginal power, marginal type I error rate, marginal FDR, and false discovery cost (FDC). Here, we demonstrate how pwrEWAS can be applied in practice using a hypothetical EWAS. In addition, we report its computational efficiency across a variety of user settings. Both under- and overpowered studies unnecessarily deplete resources and even risk failure of a study. With pwrEWAS, we provide a user-friendly tool to help researchers circumvent these risks and to assist in the design and planning of EWAS.

Availability: The web interface is written in the R statistical programming language using Shiny (RStudio Inc., 2016) and is available at https://biostats-shinyr.kumc.edu/pwrEWAS/. The R package for pwrEWAS is publicly available at GitHub (https://github.com/stefangraw/pwrEWAS).

2.2 Background

Epigenome-wide association studies (EWAS) aim to examine the relationship between epigenetic marks and exposure(s) or phenotype(s) on a genome-wide level. DNA methylation (DNAm) is the most widely studied epigenetic mechanism and involves the chemical addition of a methyl group to the 5-carbon position of cytosine in the context of cytosine-phosphate-guanine (CpG) dinucleotides. The vast majority of EWAS use microarray-based platforms for assessing DNAm, such as the Illumina Infinium HumanMethylation BeadArrays (Illumina Inc.), as these platforms provide a compromise between coverage, cost, and sample throughput (Bibikova et al., 2009; Pidsley et al., 2016). Illumina's latest methylation microarrays, the Infinium HumanMethylation450 and Infinium HumanMethylationEPIC, interrogate the methylation levels

of over 450,000 and 850,000 CpG dinucleotides, respectively. While these arrays differ in their coverage, both allow for the assessment of methylation at single-nucleotide resolution, quantified using what is referred to as the methylation β-value, an approximately continuously-distributed measure that reflects the methylation extent of a specific CpG locus; ranging from 0 (unmethylated) to 1 (methylated). Interest in studying DNAm in the context of human health and disease has been ignited by the now numerous studies that have reported altered patterns of DNAm across various human diseases (Kulis & Esteller, 2010; Robertson, 2005) and in response to environmental exposures (Martin & Fry, 2018), along with reversible nature of DNAm, which makes it a promising target for potential treatments and therapies (Yang, Lay, Han, & Jones, 2010). To detect a hypothesized difference in DNAm with adequate statistical power it is crucial to assess the required sample size. However, the complex nature of DNAm data (Saadati & Benner, 2014; Teschendorff & Relton, 2018) makes a direct power assessment challenging, as power depends on several factors: planned study sample size, array technology used to profile DNAm, tissue type used in assessing DNAm, proportion of differentially methylated CpGs and the distribution of their differences (Δ_β), and multiplicity.

The importance of formal power assessment and sample size justification in the design of research studies has been recognized and addressed in related omic fields, and motivated the development of power evaluation tools, including: "RNAseqPS" (Guo, Zhao, Li, Sheng, & Shyr, 2014), "RNASeqPowerCalculator" (Ching, Huang, & Garmire, 2014) and "PROPER" (Wu, Wang, & Wu, 2015) for RNA-Seq data, and "CaTs" (Skol, Scott, Abecasis, & Boehnke, 2006), "Statistical Power Analysis tool" (Blaise et al., 2016), "GWAPower" (Feng, Wang, Chen, & Lan, 2011), and "SurvivalGWAS_Power" (Syed, Jorgensen, & Morris, 2016) for GWAS data. However, surprisingly little attention has been given to this topic in the context of EWAS and

while there has been substantial work on the development of statistical methods and publicly available software for the preprocessing, quality control, normalization, and analysis of DNA methylation data (Li, Xie, Le Pape, & Dye, 2015; Siegmund, 2011), methods and tools for power evaluation for EWAS are lagging. Consequently, most EWAS are conducted in the absence of formal power analyses, resulting in studies that are potentially under- or overpowered (Michels et al., 2013). To our knowledge, only three studies have formally addressed the issue of power evaluation in the context of EWAS (Rakyan, Down, Balding, & Beck, 2011; Tsai & Bell, 2015; Wang, 2011). Wang et al. (Wang, 2011) simulated DNAm data for two group comparisons from uniform-normal mixture distributions with parameter settings that capture three general types of distributions often seen in methylation data (methylated, unmethylated, and partially methylated). Power was then assessed and compared for two differential methylation detection methods: proposed method by Wang et al. (Wang, 2011) and t-tests. Rakyan et al. (Rakyan et al., 2011) generated DNAm data for two group comparisons from single and mixture beta distributions in three scenarios with four effect sizes each and differences in methylation ranging from 1.25% to 14.4%. Logistic regression was then applied to assess differential methylation and power was evaluated. Finally, Tsai et al. (Tsai & Bell, 2015) simulated DNAm data for two group comparisons from nine single locus DNAm distributions, again falling into three categories: methylated, hemi-methylated and unmethylated. The expected differences in methylation ranged from 1%-60%. Differential methylation was then analyzed by t-tests and Wilcoxon rank-sum tests, and the respective power was assessed.

All three approaches utilize a limited number of single locus distributions, which result in a wide range of methylation levels of CpG sites but may lead to unrealistic data with a predefined fixed number of expected differences in methylation between two groups. This is

because individual CpGs have their own unique mean and variance depending on their genomic context and susceptibility to become methylated and vary depending on the tissue type used for methylation assessment (Lokk et al., 2014). Analogously, expected differences in CpG-specific methylation between two or more groups are expected to come from a continuous distribution instead of having predefined discrete values (Langie et al., 2017). In addition to the potential limitations above, none of the previously described methods provided accompanying software for their methodology, limiting their application within the epigenomics-research community. Therefore, there remains an outstanding need for publicly available software that addresses these limitations and enables comprehensive assessments of statistical power in the context of EWAS involving CpG-specific comparisons of DNAm.

Inspired by PROPER (Wu et al., 2015), a publicly available tool to assist researchers with power assessment in RNA-seq studies, we have developed pwrEWAS for comprehensive power evaluation in the context of case-control EWAS. In pwrEWAS, power is estimated using a semi-parametric simulation-based approach. First, DNAm data is randomly generated for each comparator group based on user-supplied information concerning the expected fraction of differentially methylated CpGs between groups and their expected effect size (Δ_β). To simulate realistic methylation data, DNAm data are generated from a beta-distribution using CpG-specific means and variances estimated from one of several different publicly available DNAm data sets, chosen to span the most common tissue-types used in EWAS. This gives the user the flexibility to select the tissue type (e.g., whole blood, peripheral blood mononuclear cells (PBMCs), etc.) that is most appropriate for the study being planned. Next, the generated data undergoes a formal differential methylation analysis, the results of which are used to estimate statistical power. In what follows, we begin by describing the statistical framework underlying pwrEWAS, followed

by its demonstration and an assessment of its run time across different user settings. We finish with a discussion of the limitations of pwrEWAS and describe future extensions.

2.3 Methods

As previously mentioned, the Illumina Infinium HumanMethylationEPIC microarray measures the methylation status of >850,000 CpGs throughout the genome. For a single CpG, DNAm is quantified via the β-value, $\beta = \frac{M}{M+U}$, where *M* and *U* are the methylated and unmethylated signal intensities, respectively. As *M* and *U* are typically assumed to be gamma-distributed random variables with equal scale parameter (Saadati & Benner, 2014), it follows that the β-value follows a beta-distribution. As such, the β-value ranges from 0 to 1 and represents the methylation extent for a specific CpG. Under ideal conditions, a β-value of zero signifies that all alleles in all cells of a sample were unmethylated at that CpG site, while a β-value of one indicates methylation throughout all alleles in all cells at that CpG site (Du et al., 2010). A common goal of EWAS is to identify CpG-specific differential methylation based on some phenotype or exposure. Formally, this involves testing the null hypothesis $H_0: \Delta_{\beta,j} = 0$, where $\Delta_{\beta,j} = \mu_j^{(1)} - \mu_j^{(2)}$ and represents the difference in mean methylation at the j^{th} CpG between two groups (e.g. cases versus controls, exposed versus unexposed, etc.), with $j = \{1, \dots, J\}$ and *J* representing the number of interrogated CpGs.

pwrEWAS is written using the R statistical programming language (http://r-project.org) and is comprised of three major steps: (1) data generation, (2) differential methylation analysis, and (3) power evaluation (Figure 4). Users are required to provide input parameters, including: tissue type to be used for methylation assessment, assumed total sample size (can be specified as a range of possible sample sizes), percentage of the total sample split into two groups (50% corresponds to a balanced study), number of CpGs to be formally tested, expected number of

differentially methylated CpGs, and the expected difference in methylation between the comparator groups (Δ_β) or alternatively, the standard deviation of these differences ($sd(\Delta_\beta)$).

To assist users with their experimental design, pwrEWAS provides estimates of statistical power as a function of the assumed sample and effect size(s). Further, it provides estimates of the marginal type I error rate, marginal FDR, false discovery cost (FDC), the distribution of simulated Δ_β's, and probabilities of identifying at least one true positive. The probability of identifying at least one true positive is beneficial in studies where either the effect or sample size is very small (e.g. pilot or explanatory studies).

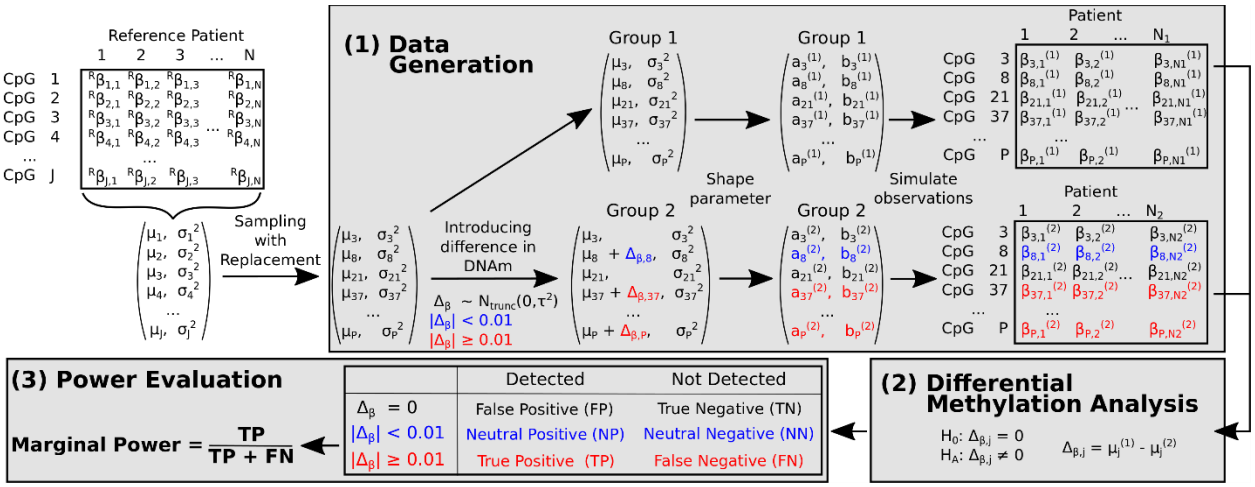


Figure 4 Workflow for pwrEWAS.

From an existing tissue-type-specific data set, J CpG-specific means and variances are estimated. Next, P CpGs are sampled with replacement from the collection of CpGs. For two groups, the mean of one group is changed by Δ_β, while the mean of the other group remains unchanged. Δ_β comes from a truncated normal distribution $N(0, \tau^2)$. These parameters are then used to simulate β-values for the two groups. A CpG with an absolute difference in mean methylation greater than a predefined detection limit (default: 0.01) is considered as truly differentially methylated. Next, the simulated data set is used to test for differential, comparing the mean methylation signatures between the two groups. A CpG is defined as "detected" if its corresponding FDR is smaller than a predefined threshold (default: 0.05). Each CpG can fall into one of six categories described in Table 1. The marginal power is calculated as the proportion of True Positives among all truly differentially methylated CpGs.

2.3.1 Data Generation

Our approach to estimating statistical power begins by leveraging publicly available DNA methylation data sets in order to simulate realistic methylation data. Data sets used for the purpose of simulation were selected to represent the most commonly used tissue types used in EWAS. To identify these tissue types, the Gene Expression Omnibus (GEO) data repository was manually scanned and tissue types were rank-ordered based on the number of GEO deposited data sets including Illumina Infinium Human Methylation BeadChip data for that tissue type. For each of the most common tissue types identified, a single representative data set was selected (Table 1). Representative data sets were selected based on a combination of the study's sample size (preference toward larger data sets), study design, and the inclusion of DNA methylation profiles for healthy, non-diseased subjects.

Table 1 Curated tissue-type specific DNAm data sets used by pwrEWAS.

Representative data sets for the most commonly used tissue types for EWAS with inclusion criteria for subjects.

Tissue Type	**Accession Number**	**Subjects within GSE-ID limited to**	**Reference**
Saliva	GSE92767		(Hong et al., 2017)
Lymphoma	GSE42372	disease state: non-HIV lymphoma	(Matsunaga et al., 2014)
Placenta	GSE62733	health state: Normal	(Kawai et al., 2015)
Liver	GSE61258	diseasestatus: Control	(Horvath et al., 2014)
Colon	GSE77718	disease state: Normal	(McInnes et al., 2017)
Blood (Adults)	GSE42861	subject: Normal	(Kular et al., 2018; Y. Liu et al., 2013)
Blood (Children)	GSE83334	age: 5 years	(Urdinguio et al., 2016)
Blood (Newborns)	GSE82273		(Markunas et al., 2016)
Cord-blood (whole blood)	GSE69176		
Cord-blood (PBMC)	GSE110128	cord blood	(Langie et al., 2018)
Adult (PBMC)	GSE67170	disease state: control	(Y. H. Zhang et al., 2018)

For each selected tissue type, CpG-specific means and variances were estimated ($\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}\beta_{i,j}$ and $\hat{\sigma}_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\beta_{i,j} - \hat{\mu}_j\right)^2$), where $\beta_{i,j}$ represents the methylation β-value for CpG $j = \{1, \dots, J\}$ in subject $i = \{1, \dots, N\}$. CpG-specific parameter estimates are then used as the basis for simulating realistic methylation data using a semi-parametric simulation strategy. First, *P* pairs of CpG-specific means and variances $\left(\hat{\mu}_j, \hat{\sigma}_j^2\right)$ are sampled with replacement from one of the tissue-type specific reference data sets (Table 1). By default, *P* is set to 100,000 CpG sites, as previous studies have suggested filtering out low-variable CpGs to offset the burden of multiplicity (Logue et al., 2017), however in principle, *P* can be set according to the user's preference (e.g., *P* = 866,836 for EWAS conducted using the EPIC array). Thus, pwrEWAS allows up- or down-scaling to any number of CpGs that the investigator plans to measure and conducted differential methylation analyses on. This is an important feature since the EPIC array is the successor to the now discontinued Infinium HumanMethylation450 array, which represents the technology used for methylation assessment of the tissue-specific reference data sets used as the basis of our simulation strategy. Of the *P* sampled CpGs, a difference in mean DNAm (Δ_β) is imposed on *K* CpGs, where $K \leq P$. The number of differentially methylated CpGs, *K*, is selected by the user and ideally motivated by a pilot study, previous literature, or expert knowledge about the effect of the phenotype(s) or exposure(s) of interest on DNA methylation. The mean methylation of *K* CpGs is shifted in one of the comparator groups by $\Delta_\beta = \left\{\Delta_{\beta,1}, \dots \Delta_{\beta,k}, \dots \Delta_{\beta,K}\right\}$, while the mean methylation in the other comparator group remains unchanged. Due to the nature of β-values and the parameter restrictions of the beta distribution ($0 \leq \mu_k \leq 1$ and $0 < \sigma_k^2 < 0.25$), $\Delta_{\beta,k}$ is bounded by $\frac{1}{2} - \mu_k \pm \sqrt{\frac{1}{4} - \sigma_k^2}$, where μ_k and σ_k^2 are CpG-specific means and variances, respectively (see **Appendix II - Effect size boundary calculation** for additional

details). Due to its boundedness, $\Delta_{\beta,k}$ is drawn from a truncated normal distribution $\left(\Delta_{\beta,k} \sim N_k(0, \tau^2)\right)$. The normal distribution was chosen based on observed differences in DNAm of differentially methylated CpGs in previously published EWAS (see Figure S2.1 in **Appendix III - Additional figures**). The standard deviation of the simulated differences τ can be provided by the user or be automatically be determined based on the user-specified target Δ_β and the expected number of differentially methylated CpGs, such that Δ_β matches the target maximal difference in mean methylation. To achieve this, an internal function simulates P $\Delta_{\beta,k}$'s (this matches the number of subsequently simulated CpGs) 100 times, while stepwise adjusting τ. The goal is to identify a standard deviation τ for the truncated normal distribution to matches the targeted maximal difference in DNAm. Therefore, τ is adjusted stepwise until the 99.99th percentile of the absolute value of simulated $\Delta_{\beta,k}$'s falls within a range around the targeted maximal difference in DNAm. The range is equal to the detection limit ($\pm$ 0.005 based on default detection limit: 0.01). Figure S2.2 (in **Appendix III - Additional figures**) shows the distribution of simulated $\Delta_{\beta,k}$'s for different effect sizes and its respective range that the 99.99th percentile of the simulated $\Delta_{\beta,k}$'s needs to fall in for τ to be accepted.

Since Δ_β is simulated from a truncated normal distribution, a certain proportion of Δ_β are within the detection limit range around zero and thus, do not exhibit a biologically meaningful difference in mean methylation. To ensure that *K* includes the number of meaningfully differential methylated CpGs (truly differentially methylated CpGs), *K* is calculated to reflect the user-supplied target number of differentially methylated CpGs ($K = \frac{1}{Percentage\ of\ truly\ DM\ CpGs} * Target\ number\ of\ DM\ CpGs$). This results in *K* CpGs with changed means ($\Delta_{\beta,k} \neq 0$) and $P - K$ CpGs with unchanged means ($\Delta_{\beta,k} = 0$) between the two comparator groups. Variances across

all *P* CpGs remain unchanged in both comparator groups, that is, comparator groups are assumed to have the same CpG-specific variances. Next, the means and variances of both comparator groups are used to calculate CpG-specific shape parameters for the beta-distribution: $a_j = \mu_j^2\left(\frac{1-\mu_j}{\sigma_j^2} - \frac{1}{\mu_j}\right)$ and $b_j = a_j\left(\frac{1}{\mu_j} - 1\right)$ (see **Appendix II - Effect size boundary calculation**). The two comparator group specific matrices (*P* x *2*) containing the CpG-specific shape parameters are then used to generate N_1 and N_2 beta-distributed observations for each CpG, for both comparator groups respectively, resulting in two matrices (*P* x N_1 and *P* x N_2) of β-values, which are subsequently used for the differential methylation analysis.

Simulated CpGs fall into one of three categories: (1) not differentially methylated ($\Delta_{\beta,k} = 0$), (2) differentially methylated with negligible difference ($|\Delta_{\beta,k}| < 0.01$), and (3) truly differentially methylated ($|\Delta_{\beta,k}| \geq 0.01$). The threshold of 0.01 was chosen according to the detection limit of DNAm arrays (Teschendorff & Relton, 2018), but can be modified by the user.

2.3.2 Differential Methylation Detection

Following data generation, differential methylation analyses are carried out using one of several established parametric and nonparametric approaches, including: limma (Ritchie et al., 2015), CpGassoc (Barfield, Kilaru, Smith, & Conneely, 2012), t-test, or a Wilcoxon rank-sum test. In the first three of the above methods, simulated β-values are first transformed to methylation *M*-values using the logit-transformation ($M = \log_2\left(\frac{\beta}{1-\beta}\right)$) due to their assumption of normality (Du et al., 2010; Wilhelm-Benartzi et al., 2013). Each method reports CpG-specific p-values, which are multiplicity adjusted using the Benjamini and Hochberg method (Benjamini & Hochberg, 1995) to control the False Discovery Rate (FDR).

2.3.3 Power Assessment

Tested CpGs fall into one of six categories: (1) TP (True Positive): detected CpGs with meaningful difference in mean DNAm, (2) NP (Neutral Positive): detected CpGs with negligible difference in mean DNAm, (3) FP (False Positive): detected CpG with no difference in mean DNAm, (4) TN (True Negative): undetected CpGs with no difference in mean DNAm, (5) NN (Neutral Negative): undetected CpGs with negligible difference in mean DNAm, and (6) FN (False Negative): undetected CpGs with meaningful difference in mean DNAm (Table 2).

Table 2 Differential methylation detection and terminology.

Each CpG can fall into one of six following categories: False Positive (FP; detected CpG with no simulated difference in mean methylation); Neutral Positive (NP; detected CpG with negotiable simulated difference in mean methylation); True Positive (TP; detected CpG with meaningful simulated difference in mean methylation); True Negative (TN; not detected CpG with no simulated difference in mean methylation); Neutral Negative (NN; not detected CpG with negotiable simulated difference in mean methylation); False Negative (FN; not detected CpG with meaningful simulated difference in mean methylation).

	Differentially Methylated	**Truly Differentially Methylated**	**Detected**	**Not Detected**
$\boldsymbol{\Delta_k = 0}$	No	No	False Positive (FP)	True Negative (TN)
$\boldsymbol{\lvert\Delta_k\rvert < 0.01}$	Yes	No	Neutral Positive (NP)	Neutral Negative (NN)
$\boldsymbol{\lvert\Delta_k\rvert \geq 0.01}$	Yes	Yes	True Positive (TP)	False Negative (FN)

Since it can be argued that CpGs with a negligible $\Delta_{\beta,k}$ are not biologically meaningful, we calculate the empirical marginal power, defined by Wu et al. (Wu et al., 2015) as the proportion of truly differentially methylated CpGs detected at the specified FDR threshold, $\frac{TP}{TP+FN}$ (Table 2). Further, even though failing to discover differentially methylated CpGs

represents a type II error, failing to detect CpGs with a negligible $\Delta_{\beta,k}$ can be disregarded (NN) due to their likely unimportance. Additionally, as identifying CpGs with a negligible $\Delta_{\beta,k}$ (NP) is not as crucial as identifying CpGs with a biologically meaningful $\Delta_{\beta,k}$ (TP), we also report the false discovery cost ($FDC = \frac{FP}{TP}$) (Wu et al., 2015).

For each of the assumed sample and effect sizes we report the following metrics, averaged across simulations to obtain reliable estimates:

Empirical classical power: The ratio of correctly detected CpGs and all differentially methylated CpGs

$$classicalPower = \frac{NP + TP}{NP + NN + TP + FN}$$

Empirical marginal power: The ratio of correctly detected CpGs with biologically meaningful differences and all differentially methylated CpGs with biologically meaningful differences (excluding Neutral Positives and Neutral Negative with negligible differences):

$$marPower = \frac{TP}{TP + FN}$$

Empirical marginal Type I Error: The ratio of wrongly detected CpGs and all CpGs with no difference

$$marTypeI = \frac{FP}{FP + TN}$$

Empirical False Discovery Rate (FDR): The ratio of wrongly detected CpGs and all detected CpGs

$$FDR = \frac{FP}{FP + NP + TP}$$

Empirical False Discovery cost (FDC): The ratio of wrongly detected CpGs and correctly detected CpGs:

$$FDC = \frac{FP}{TP}$$

2.3.4 Visualization

The pwrEWAS package contains two functions that can be used to visualize the results ("pwrEWAS_powerPlot" and "pwrEWAS_deltaDensity"). "pwrEWAS_powerPlot" displays the estimated power as a function of sample size with error bars (2.5th and 97.5th percentile calculated across simulations). Power across different target $\Delta_\beta{}'$s as a function of sample size is differentiated by different colors (Figure 5, Box 4). "pwrEWAS_deltaDensity" illustrates the distribution of simulated $\Delta_{\beta,k}{}'$s for different target $\Delta_\beta{}'$s as density plots (Figure 5, Box 7). Densities for different target $\Delta_\beta{}'$s are color-coded as well and match the colors of the power curve ("pwrEWAS_powerPlot").

pwrEWAS

pwrEWAS is a computationally efficient tool to estimate power in EWAS as a function of sample and effect size for two-group comparisons of DNAm (e.g., case vs control, exposed vs non-exposed, etc.). Detailed description of in-/outputs, instructions and an example, as well as interpretations of the example results are provided in the following vignette: pwrEWAS vignette

3

Authors: Stefan Graw, Devin Koestler
Department of Biostatistics, University of Kansas School of Medicine

1

Tissue Type
Blood adult

Minimum total sample size
20

Maximum total sample size
260

Sample size increments
40

Samples rate for group 1
0.5

Number of CpGs tested
100000

Target number of DM CpGs
2500

Target max Δ | SD(Δ)

Target maximal difference in DNAm (comma delimited)
0.02, 0.10, 0.15, 0.20

Target FDR
0.05

2

☑ Advanced settings

Detection Limit
0.01

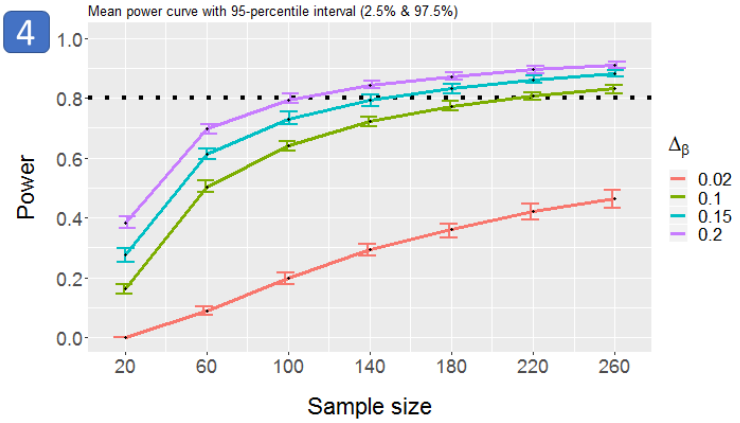
Method for DM analysis
limma

Number of simulated data sets
50

Threads
4

Go!

4



5

Power

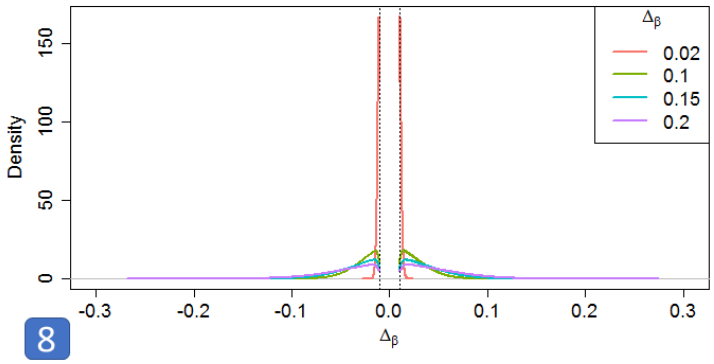
N \ Δβ	0.02	0.1	0.15	0.2
20	0	0.16	0.28	0.38
60	0.09	0.5	0.61	0.7
100	0.2	0.64	0.73	0.79
140	0.29	0.72	0.79	0.84
180	0.36	0.77	0.83	0.87
220	0.42	0.81	0.86	0.9
260	0.46	0.83	0.88	0.91

6

P(#TP≥1)

N \ Δβ	0.02	0.1	0.15	0.2
20	0.36	1	1	1
60	1	1	1	1
100	1	1	1	1
140	1	1	1	1
180	1	1	1	1
220	1	1	1	1
260	1	1	1	1

7



8

```
Tissue type = Blood adult
Minimum total sample size = 20
Maximum total sample size = 260
Sample size increments = 40
Percentage samples in group 1 = 0.5
Number of CpGs to be tested = 100000
Target number of DM CpGs = 2500
'Target max Delta' was selected
Target maximal difference in DNAm (comma delimited) = 0.02, 0.10, 0.15, 0.20
Target FDR = 0.05
Detection Limit = 0.01
Method for DM analysis = limma
Number of simulated data sets = 50
Threads = 4
Run time = 49.2 mins
```

Figure 5 pwrEWAS Shiny User-Interface.

(1) User-specific inputs; (2) Advanced input settings to optimize run time; (3) Link to vignette for detailed description of inputs and outputs, instructions and an example including interpretations of the example results; (4) Power curve as a function of sample size by effect size (Δ_β); (5) Estimated power average over simulation by sample size and effect size (Δ_β); (6) Probability of detection at least one true positive; (7) Distribution of simulated differences in DNAm (Δ_β) for different target Δ_β's; (8) Log of input parameter and run time.

2.4 Results

Consider a hypothetical study that aims to understand the relationship between electronic cigarettes (e-cigarette) and DNAm derived from adult blood. The use of e-cigarettes has increased dramatically over the last decade, especially among young adults (H. Chen et al., 2018). There exists a common perception in the population, including pregnant women and women in child-bearing age, that e-cigarettes are less harmful than smoking tobacco cigarettes (Nguyen et al., 2018). Although, studies have reported the presence of toxic components in e-cigarette aerosol (H. Chen et al., 2018), there presently exists no study investigating the relationship between e-cigarette and DNAm derived from adult human blood. As the effect of e-cigarette usage on DNAm is presently unknown, but is of interest in this hypothetical study, we will use the previously reported effects of tobacco smoke on blood-derived DNAm as an upper limit for the effect of e-cigarette usage on DNAm. Previous studies analyzing the effect of smoking tobacco cigarettes on blood-derived patterns of DNA methylation have reported CpG-specific differences up to 24% between smokers and non-smokers, with a wide range of CpGs (724 - 18,760) declared as significantly differentially methylated (FDR ≤ 0.05) (Ambatipudi et al., 2016; Joehanes et al., 2016; Zeilinger et al., 2013). Hence, we want to investigate the number of subjects required to detect DNAm differences in 2,500 CpGs (selected to be within the range

of the number of significantly differently methylated CpGs reported between smokers and non-smokers in previous reports) with 80% power for three reasonable effect sizes ($\Delta_\beta = \{0.10, 0.15, 0.20\}$ and one deliberately small effect size $\Delta_\beta = 0.02$, representing differences in DNAm up to $\sim$2%, $\sim$10%, $\sim$15% and $\sim$20%). To cover a wide range of total sample sizes, we analyzed total sample sizes ranging from 20 to 260 individuals with increments of 40 and equal allocation between e-cigarette users and non-users, while keeping the remaining default parameters of pwrEWAS intact:

- Tissue type: Blood adult
- Minimum total sample size: 20
- Minimum total sample size: 260
- Sample size increments: 40
- Samples rate for group 1: 0.50
- Number of CpGs tested: 100000
- Target number of DM CpGs: 2500
- Select 'Target max Δ'
- Target maximal difference in DNAm: 0.02, 0.10, 0.15, 0.20
- Target FDR: 0.05
- Detection Limit: 0.01
- Method for DM analysis: limma
- Number of simulated data sets: 50
- Threads: 4

The results of this power analysis can be found in Figure 5. To detect differences up to 10%, 15% and 20% in CpG-specific methylation across 2,500 CpGs between e-cigarette users

and non-users with at least 80% power, we would need about 220, 180 and 140 total subjects, respectively. As expected, 80% power was not achieved for a difference in DNAm $\leq 2\%$ for the selected total sample size range. However, it can be observed for this target differences of 2% that the probability of detecting at least one CpG out of the 2,500 differentially methylated CpGs is about 36% for 20 total patients and virtually 100% for 60 and more total patients. Because there exists no literature on the magnitude of expected differences in DNAm, a pilot study would be helpful in this hypothetical situation to narrow the range of expected differences to more precisely identify the required sample size to achieve 80% power.

To evaluate this broad range of sample and effect sizes of this theoretical experiment, pwrEWAS required ~49min in total. In general, the computational complexity of pwrEWAS depends on four major components: (1) assumed number and magnitude of sample size(s), (2) number of target Δ_β's (effect sizes), (3) number of CpGs tested, and (4) number of simulated data sets. To enhance the computational efficiency, pwrEWAS allows users to process simulations in parallel. While (1) and (2) are usually dictated by the study to be conducted, (3) and (4) can be modified to either increase the precision of power estimates (increased run time) or reduce the computational burden (decreased precision of estimates). The run time of pwrEWAS for different combinations of sample sizes and effect sizes are provided in Table 3.

Table 3 Run time of pwrEWAS.

Run time of pwrEWAS for different combinations of sample sizes and effect sizes. In all scenarios presented the number of tested CpGs was assumed to be 100,000, number of simulated data sets was 50, and the method to perform the differential methylation analysis as limma. A total of 6 clusters/threads were used.

Total sample sizes	**Effect sizes (Δ_β)**		
	0.1	**0.1, 0.2**	**0.1, 0.3, 0.5**
10	2min 21sec	3min 11sec	3min 50sec
100	6min 22sec	7min 39sec	8min 33sec
500	24min 43sec	27min 36sec	29min 22sec
10-100 (increments of 10)	9min 40sec	16min 34sec	23min 44sec
300-500 (increments of 100)	27min 58sec	30min 01sec	52min 00sec

As the number of simulated data sets is one of the major components (e.g., item (4), above) affecting the run time of pwrEWAS, it is important to identify a default value that offers a reasonable tradeoff between run time and precision of power estimates. To this end, the variance of power estimates was assessed for a range of simulated data sets (5-100), each repeated 100 times, while keeping the remaining parameters unchanged (Figure 6A). We ultimately determined the default value for the number of simulated data sets to be 50, as it appears that simulating additional data sets reduces the variance of power estimates only marginally (Figure 6B).

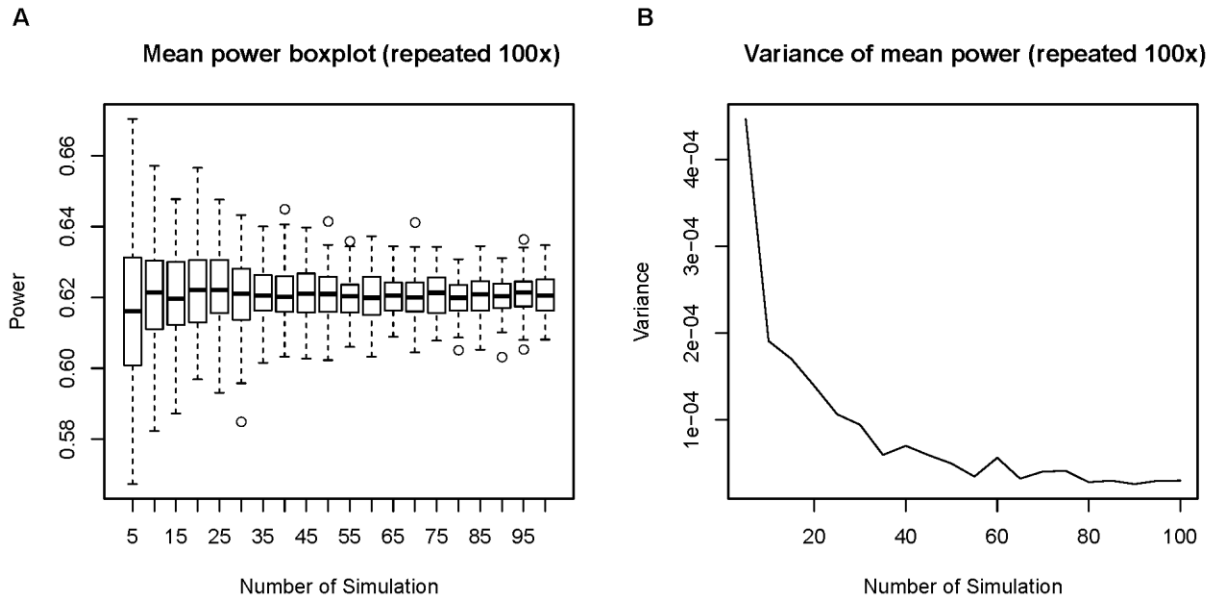


Figure 6 Empirical assessment of the number of simulations.

To assess the number of simulated data sets (number of simulations) required to obtain consistent results for power, pwrEWAS was run for a variety of number of simulations (5-100 simulations), each 100 times and each with the same remaining input parameters. (A) shows the distribution of power estimates for 100 runs within each of the assumed number of simulations. (B) visualizes the variance of power estimates for each of the assumed number of simulations. Given the relative stability of variance estimates beyond 50 simulations, 50 was selected as the default value for the number of simulations in pwrEWAS.

The pwrEWAS package is accompanied by a vignette (**Appendix IV - Vignette**), which provides a more detailed description of input and output, instructions for the usage, an example, and interpretations of the example results. In addition, a user-friendly R-Shiny point-and-click interface has been developed (Figure 5) for researchers that are unfamiliar or less comfortable with the R environment.

2.5 Discussion

In our hypothetical study on the effect of e-cigarette usage on patterns of blood-derived DNAm, we found that 140-220 total subjects would be needed, depending on the expected effect size. However, these results should be treated with a certain level of caution and considered to be more of a guideline than an exact prescription. Due to computational, memory and storage burden, and simplicity considerations, pwrEWAS involves the random generation methylation β-values independently across CpGs, which might not hold in real data given previous reports of local correlation in DNAm of nearby CpG sites (W. W. Zhang, Spector, Deloukas, Bell, & Engelhardt, 2015). Additionally, pwrEWAS assumes CpG-specific homoscedasticity between both comparator groups, that is CpG-specific variances are assumed to be identical between both groups. However, CpG-specific variances have been reported to change depending on exposure(s) and phenotype(s) (Hansen et al., 2011; Teschendorff et al., 2012). Violations of CpG-specific homoscedasticity can result in inflated estimates of statistical power and produce overly optimistic sample sizes, however identifying the magnitude of changes in variances depending on exposure(s) and phenotype(s) in advance of the study can be very challenging. Further, the expected difference in DNAm between both groups (Δ_β) is assumed to come from a truncated normal. This assumption seems to hold, at least approximately, based on observed distributions of differences in DNAm across a variety of studies. Additional limitations of pwrEWAS include: two group comparison, selection of methods for differential methylation analysis, and selection of tissue types specific reference data.

Despite the above limitations, pwrEWAS is to our knowledge the first publicly available tool to formally address the issue of power evaluation in the context of EWAS. Further opportunities for the extension of pwrEWAS include the implementation of additional methods

for differential methylation analysis (e.g., linear regression for continuous phenotype(s)/exposure(s), Cox-proportional hazards models or relevant models for handling time-to-event outcomes, etc.), allowing multiple group comparisons, providing the opportunity for researcher to upload different reference data (tissue type(s) specific to their study), and addressing the potential change of CpG dispersion due to phenotype(s) and/or exposure(s).

2.6 Conclusion

When designing an EWAS, consideration of statistical power should play a central role in selecting the appropriate sample size to address the question(s) of interest. Under- and overpowered studies waste resources and even risk failure of the study. With pwrEWAS we present a user-friendly power evaluation tool with the goal of helping researchers in the design and planning of their EWAS.

2.7 Contribution

In this project I conceived the idea and framework of pwrEWAS, implemented pwrEWAS (including its visualizations and shiny user-interface), created the R package "pwrEWAS", demonstrated pwrEWAS's application in a hypothetical study, assisted in the process of creating reference data sets and wrote the manuscript and vignette.

3 Chapter 3: Selection probability optimization for feature selection (SPOFS): A feature selection strategy for biomarker identification

3.1 Abstract

Biomarkers offer great promise in improving disease diagnosis, prognosis, and the choice of treatment. In identifying useful biomarkers among a set of many candidate biomarkers, feature selection techniques have proven to be effective methods, especially in high-dimensional omic data sets. However, the high dimensionality of DNA methylation array data along with its unique characteristics render the identification of useful biomarkers in such studies challenging.

Motivated by the goal of identifying blood-based DNA methylation biomarkers in a nested case-control study of lung cancer risk and the subsequent development of a prediction model based on such biomarkers, we present a heuristic feature selection approach called Selection Probability Optimization for Feature Selection (SPOFS). SPOFS is designed to identify optimal subset(s) of epigenetic features from among a vast pool of such features, such that the resultant subset(s) optimize the ability to predict some outcome or response. As an initial step, SPOFS involves filtering the data to a manageable number of features. Next, sets of features are iteratively selected and the contribution of each selected feature to the prediction performance is assessed. The selection probability of each evaluated feature is then modified based on its individual performance, in which the selection probabilities are increased for highly predictive features and reduced for less predictive features. The final selected model consists of K epigenetic features with the highest selection probabilities or the model that achieved the maximum prediction performance during the course of the previously described iterative procedure.

We evaluated SPOFS using both data from a nested-case control (NCC) study of lung cancer risk and in simulated data and benchmarked its performance against Lasso (Least Absolute Shrinkage and Selection Operator). The objective of the lung cancer risk NCC study was to build prediction model(s) for predicting lung cancer risk using pre-diagnostic blood-derived DNA methylation biomarkers. In simulated data, we demonstrate that SPOFS outperforms Lasso in scenarios involving strong correlation between risk biomarkers, and when those biomarkers are associated with small effects relative to the outcome. Comparable performance was observed between SPOFS and Lasso in scenarios of weak to moderate correlation between biomarkers and when the effect of biomarkers was moderate to large. In the analysis of the lung cancer risk NCC study, we establish that neither SPOFS nor Lasso was able to identify a set of biomarkers that resulted in satisfying prediction accuracy. While these results may suggest that blood-derived DNA methylation has limited signal for predicting lung cancer risk, the applicability of SPOFS to real data sets remains an outstanding need. While this work provides initial support for SPOFS as a competitive feature selection method, there remain numerous opportunities for future development, enhancement, and evaluation of this approach.

3.2 Background

Biological markers or biomarkers were first defined in 1998 by the National Institutes of Health Biomarkers Definitions Working Group as "*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention*" (Biomarkers Definitions Working Group, 2001), and represent powerful tools for decision-making processes in disease monitoring and treatment selection. For any newly proposed biomarker, it is critical to evaluate its accuracy for predicting future events (e.g. patient survival, disease onset, or recurrence) in prospective

studies. Many well-known studies, including the Carotene and Retinol Efficacy Trial (CARET) (Goodman et al., 2004), recognized this need and proactively stored biospecimens collected from study participants during their enrollment in the trial. However, the assessment of biomarkers can be expensive and time-consuming to collect, rendering their assessment in the full cohort impractical and infeasible. To circumvent these limitations, nested case-control (NCC) study designs involve the assessment of the biomarker(s) of interest in cases and a subset of controls, matched to individual cases based on their clinical characteristics, rather than assessing the biomarker across the entire cohort. While the NCC design is particularly advantageous when the prevalence of a disease is rare, biomarker assessment is costly to perform, or if the test is invasive (Biesheuvel et al., 2008), the sampling strategy of NCC designs can introduce methodological challenges in the statistical analysis and complicate the identification of useful biomarkers in high-dimensional settings.

One of the foremost analytical challenges associated with NCC studies is the incorporation of case-control matching in the statistical analysis. Conditional logistic regression is an extension of the logistic regression, where matching is accounted for by introducing a stratification term unique to each matched case-control group (Liddell, Mcdonald, & Thomas, 1977). While conditional logistic regression is the most flexible and general procedure for matched data, this approach does not explicitly facilitate the incorporation of time-to-event information. Information on the length of time between biospecimen collection (e.g. time-point at which the biomarker was assessed) and diagnosis is key criteria used in defining the risk set for each case, and thus matching. This time difference is not explicitly modeled in the conditional logistic regression model. Inverse Probability Weighting (IPW) in the context of Cox proportional-hazards models is an alternative approach, where individuals are assigned weights

inversely proportional to their inclusion probabilities of being sampled in the NCC study (Samuelsen, 1997). Even though Samuelsen's IPW approach has been shown to be more efficient than conditional logistic regression (Kim, 2013), conditional logistic regression remains as the most commonly used methodology to analyze data from a NCC study. For complex data, the statistical method of feature selection is important in the process of extracting knowledge and part of ongoing research (Lee & Krischer, 2017).

Many feature selection approaches have been proposed in the context of conditional logistic regression and have been categorized into three groups: (1) relevant features based on test statistics of original and modified pair t-tests, often followed by a classification approach, (2) feature selection algorithms within the conditional logistic regression framework, and (3) boosting strategies that address classification problems with matched case-control responses (Liang, Ma, Yang, Wang, & Ma, 2018). The best performance and accuracy was achieved by boosting strategies, while t-test methods performed poorest. However, the computational burden and run time of the boosting strategies pose a challenge as the number of features increases. While several approaches have been proposed in the context of Cox proportional-hazards models to address the issue of feature selection, only a handful of them are able to manage high-dimensional data, where the number of covariates (P) is greater than the sample size (N). Grace and Li provide a selective review of feature selection methods for high-dimensional data, and underline the importance of their continuous development (Grace & Li, 2017).

With growing high-dimensional data sets, the number of possible feature combinations increases rapidly. As a result, computationally efficient strategies for identifying optimal combinations of features is crucial. Inspired by IDOL (Koestler et al., 2016), a dynamic algorithm for Identifying Optimal Libraries in the context of cell mixture deconvolution using

DNA methylation (DNAm) data, we present Selection Probability Optimization for Feature Selection (SPOFS) for identifying predictive features in the context of inverse probability weighted Cox regression models. SPOFS is a heuristic algorithm where features are iteratively selected randomly, each feature's importance is evaluated within the sampled subset of features, and their selection probability for subsequent iterations is modified proportional to contribution to prediction performance. Adapting this method from IDOL was motivated by a NCC study from the Carotene and Retinol Efficacy Trial (CARET) whose goal involved the identification of a set of blood-derived DNAm biomarkers that collectively predict lung cancer risk. To benchmark our proposed method, we compared its performance to Lasso (Tibshirani, 1996) across different simulation scenarios and in the lung cancer NCC study.

In the following sections, we first orient the reader by providing general notation involving time-to-event analysis with the Cox proportional-hazards model and introduce the concept of IPW for NCC studies. Next, we provide a comprehensive description of SPOFS, followed by a high-level explanation of Lasso. We then describe the details of our simulation study to assess the performance of SPOFS. Finally, we report the results obtained from SPOFS and Lasso, applied to the lung cancer NCC study, as well as to simulated data across different scenarios. We conclude by discussing our findings and describe future directions for this work.

3.3 Materials and Methods

3.3.1 General notation and IPW Cox regression modeling

Suppose we have a cohort of N patients with P covariates, where $P > N$. Let t_i and c_i denote the event and censoring time, respectively, for patient i (e.g. time from blood-draw to diagnosis or censoring), where $i = \{1, \dots, N\}$. Let $Y_i = \min(t_i, c_i)$ and event indicator $\delta_i = I(t_i \leq c_i)$, where $I(\cdot)$ represents an indicator function. In general, we assume that t_i and c_i are

independent. Further, let $\boldsymbol{X_i} = (X_{i1}, \dots, X_{iP})^T$ denote a vector with P biomarkers of interest and an additional vector $\boldsymbol{Z_i} = (Z_{i1}, \dots, Z_{iV})^T$with V matching variables for the i^{th} patient.

In the NCC design setting, at some time point t_c, there exist n individuals who have experienced the event at time t_i, such that $t_i < t_c$. These individuals are selected as "cases". In addition, for each case i a random sample of potential controls is selected from the case specific risk set, such that $R(t_i) = \{l: Y_l \geq t_i\}$ and $l = \{1, \dots, n_l\}$. Without loss of generality, we assume that the number of matched controls for each case is equal to one (e.g., 1:1 matching). It is important to note that case A can operate as a potential control for another case B, provided that the time at risk for case A is greater than the time at risk of case B ($t_A \geq t_B$). Here, we utilize the Cox proportional hazard model defined as:

$$h_i(t|X_i, Z_i) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X_i} + \boldsymbol{\gamma}^T\boldsymbol{Z_i})$$

where $h_0(t)$ represents the baseline hazard and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the log-hazard ratios for $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$, respectively. As the method of inverse probability weighting (IPW) allows for breaking the matching in NCC studies, it is important to adjust for matching variables ($\boldsymbol{Z_i}$) (Stoer & Samuelsen, 2013). However, the estimation of parameters is not trivial due to the matching, and cases and controls are only used at the event time of a particular case (Stoer & Samuelsen, 2013). An estimation procedure based on the following weighted partial likelihood was proposed by Samuelsen (Samuelsen, 1997):

$$L(\beta, \gamma) = \prod_i \frac{\exp(\boldsymbol{\beta^T X_i} + \boldsymbol{\gamma^T Z_i})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta^T X_l} + \boldsymbol{\gamma^T Z_l})\, w_l} \qquad (3.3.1.1)$$

The partial likelihood is a product over all i cases, divided by the sum of the risk set R at time t_i, defined as all matching individuals at risk for that case. The weight w_l represents the inverse probability of an individual l having been sampled and is equal to 1 for all cases. For

simplicity, it is assumed here that all covariates are time invariant, although time-varying covariates are possible (Stoer & Samuelsen, 2013).

IPW aims to adjust for bias in sampling from the full cohort, specifically the sampling bias with respect to the proportions of cases and controls, but also sampling bias with respect to matching variables introduced by additional matching. The idea is to let each control represent the complete risk set $R(t_i)$ by assigning them a weight greater than 1. As the risk set size increases, the probability of an individual to be selected as a control from this risk set decreases, and hence its weight increases. Assuming the selected controls are representative of the respective risk set, analyses and inferences can be conducted "as if the data was from a cohort study", by introducing the inverse probability weights and utilizing a weighted Cox-regression model (Stoer & Samuelsen, 2013, p. 6).

3.3.2 Weight estimation

The weights w_l for each selected control *l* in equation (3.3.1.1) must be estimated from the full cohort using the matching variables considered in the NCC study. Different types of estimators have been proposed including: Kaplan-Meier type of weights (Cai & Zheng, 2012; Salim, Hultman, Sparen, & Reilly, 2009; Samuelsen, 1997), model-based, logistic regression type weights (Mark & Katki, 2006; Saarela, Kulathinal, Arjas, & Laara, 2008; Samuelsen, Aring;nestad, & Skrondal, 2007; Stoer & Samuelsen, 2013), and local averaging weights (K. N. Chen, 2001). Here, we used the function "KMprob" in the R package *multipleNCC* (Stoer & Samuelsen, 2016) to estimate weights of Kaplan-Meier type, defined as follows:

$$\frac{1}{w_l} = 1 - \prod_{t_i < t_l} \left\{ 1 - \frac{1}{n_i(t_i) - 1} I(\text{Control } l \text{ meets matching criteria for case } i) \right\}$$

Here $n_i(t_i)$ is the number of individuals at risk at time t_i that meet the matching criteria of case i and $I(\cdot)$ is an indicator function restricting the product to only include individuals that meet the matching criteria.

3.3.3 Selection probability estimation for feature selection (SPOFS)

Our goal involves the identification of an optimal subset of epigenetic features from among a vast pool of such features, such that the resultant subset(s) optimizes the ability to predict lung cancer risk beyond the prediction accuracy achieved by models based on clinical features only. In what follows, we provide a detailed description of each step of the SPOFS algorithm.

Initial filtering step: **Epigenetic feature assessment and filtering**

1) For each epigenetic feature, a weighted Cox regression model including this epigenetic feature, as well as all matching features (clinical features), is fit to the data using 10-fold cross-validation (CV). That is, a model is repeatedly fit to 9/10 of the data and evaluated on the 10^{th} held-back piece, until each of the 10 pieces was held back once. The results of each fold are then averaged. Here, each epigenetic feature's importance is evaluated based on its concordance index (*c*-index). The *c*-index is a popular metric to evaluate prediction performance in time-to-event analyses. It is defined as the probability that for a pair of randomly chosen individuals, the individual with the higher risk prediction experiences the event first (see **3.3.4 Concordance Index**).
2) Epigenetic features are then rank-ordered based on their average c-index and filtered such that a set Q of best performing epigenetic features remains. Here Q is defined as the P epigenetic features with the largest average *c*-index.

SPOFS: **Scanning for optimal set of epigenetic features by updating their selection probability based on performance**

1) At each iteration *m*, *K* epigenetic features are randomly selected from $\mathcal{Q}$ with probability $\pi_p^{(m)}$, where $p = \{1,2, \dots, P\}$. At iteration $m = 0$, the selection probability for each epigenetic feature contained in $\mathcal{Q}$ is specified to be equal, that is, $\pi_p^{(0)} = \frac{1}{P}$.
2) Let the subset of sampled epigenetic features be defined as $\mathcal{Q}^{(m)} \subset \mathcal{Q}$ at iteration *m.*

The following steps are performed using 10-fold CV.

3) Fit a full model (including all *K* sampled epigenetic features) by applying the 2-step modeling procedure proposed by Thompson et al. (Thompson, Christensen, & Marsit, 2018), where an initial model is built using only the epigenetic features, in order to generate an epigenetic risk score. These epigenetic risk scores are then used to build a final model in conjunction with clinical features.
    a) An individual specific epigenetic risk score $\nu_i^{(m)}$ is calculated by fitting a weighted Cox regression model using only the epigenetic features contained in $\mathcal{Q}^{(m)}$.
    b) A subsequent weighted Cox regression model is then trained using the epigenetic risk scores $\nu^{(m)} = \left\{\nu_1^{(m)}, \dots, \nu_N^{(m)}\right\}$ and matching features (clinical features) as terms in the model.
    c) The performance of the 2-step full model is assessed at iteration *m* by calculating its *c*-index $c_F^{(m)}$, where *F* refers to the full model.

4) Fit *K* reduced models to a reduced set of epigenetic features $Q_{-k}^{(m)}$, where the k^{th} epigenetic feature is omitted from $Q^{(m)}$ and $k = \{1, \dots, K\}$, by applying the same 2-step modeling procedure described above:
    a) Individual specific epigenetic risk scores $v_{i,-k}^{(m)}$ are calculated by fitting *K* weighted Cox regression model to the reduced set of epigenetic features $Q_{-k}^{(m)}$.
    b) Subsequent weighted Cox regression models are then trained using the epigenetic risk scores $v_{-k}^{(m)} = \left\{v_{1,-k}^{(m)}, \dots, v_{N,-k}^{(m)}\right\}$ and the matching variables as terms in the model.
    c) Performance of the *K* reduced models are assessed by calculating their *c*-indices $c_{R_k}^{(m)}$ at iteration *m*.
5) The contribution of each epigenetic feature in $Q^{(m)}$ is assessed by calculating the difference between the *c*-indices of the full and reduced models: $\Delta_k^{(m)} = c_F^{(m)} - c_{R_k}^{(m)}$, where $\Delta_k^{(m)}$ can fall into one of the following three scenarios:
    a) $0 < \Delta_k^{(m)} \leq 1$: Removing the k^{th} epigenetic feature leads to poorer overall prediction performance, and therefore the selection probability of k^{th} epigenetic feature should be increased in subsequent iterations. As $\Delta_k^{(m)}$ increases, we expect increasing contribution of k^{th} epigenetic feature to overall prediction performance.
    b) $\Delta_k^{(m)} = 0$: Removing the k^{th} epigenetic feature results in no change in performance, and therefore the selection probability of k^{th} epigenetic feature should remain the same.

c) $-1 \leq \Delta_k^{(m)} < 0$: A model without the k^{th} epigenetic feature performs better compared to a model containing it, and therefore the selection probability of k^{th} epigenetic feature should be decreased in subsequent iterations.

End of 10-fold CV.

6) Based on the contribution to the prediction performance of each selected epigenetic feature, represented by $\Delta_k^{(m)}$, its probability of being selected in subsequent iteration is adjusted, such that epigenetic features with greater performance contribution are more likely to be chosen. To achieve this, the following selection probability properties are defined:

   a) Every epigenetic feature *p* has the same initial probability of selection: $\pi_p^{(0)} = \frac{1}{P}$

   b) The probability of a specific epigenetic feature to be selected must be between 0 and 1: $0 \geq \pi_p^{(m)} \geq 1$

   c) The sum over the selection probability of all epigenetic features is equal to 1: $\sum_{p=1}^{P} \pi_p^{(m)} = 1$

   d) The sum over the selection probability of the *K* selected epigenetic features is identical before and after modifying their selection probability: $\sum_{k=1}^{K} \pi_k^{(m-1)} = \sum_{k=1}^{K} \pi_k^{(m)}$

7) The selection probability of each selected epigenetic feature *k* at iteration *m* is a function of its selection probability at iteration $m-1$ and a weighting function based on $\Delta_k^{(m)}$. This weighted selection probability is then scaled by the sum of weighted selection probabilities of all selected epigenetic feature as well as the sum of their selection probabilities in iteration $m-1$:

$$\pi_k^{(m)} = \frac{\pi_k^{(m-1)} \cdot f\left(\Delta_k^{(m)}\right)}{\sum_{k'=1}^{K} \pi_{k'}^{(m-1)} \cdot f\left(\Delta_{k'}^{(m)}\right)} * \sum_{k'=1}^{K} \pi_{k'}^{(m-1)}$$

Note: The weighting function $f(\cdot)$ can be any non-decreasing function, supporting values between -1 and 1. The choice of function is driven by the distribution of $\Delta_k^{(m)}$ and dictates the degree to which the selection probabilities are modified. Possible functions that were considered are listed in **Appendix V - List of potential weighting functions**.

8) Repeat main steps 1-7, while recording: full model at iteration *m*, the set of selected epigenetic features $\mathcal{Q}^{(m)}$ and prediction performance (10-fold cross-validated $c_F^{(m)}$).

3.3.4 Concordance Index

The concordance index (c-index) is a frequently used metric to evaluate the prediction performance of models for time-to-event outcomes. It was first proposed by Mann & Whitney in 1947 as a test of whether one of two random variables is stochastically larger than the other (Mann & Whitney, 1947). More formally, let S_1 and S_2 be continuous independent random variables (e.g., linear predictors). The c-index is defined as:

$$c = P(S_1 > S_2)$$

In the context of a time-to-event analysis, the c-index is defined as the probability that for a pair of randomly chosen individuals, the individual with the higher risk prediction (e.g., larger value of the linear predictor) experiences the event first. A c-index > 0.50 indicates a good prediction performance of a model, while a c-index of 0.50 implies that a model has no predictive ability. The c-index was calculated using the function “concordance.index” contained in the R package *survcomp*, which calculates the c-index by evaluating the risk scores of pairs of

randomly chosen samples (Haibe-Kains, Desmedt, Sotiriou, & Bontempi, 2008; Schroder, Culhane, Quackenbush, & Haibe-Kains, 2011).

3.3.5 Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) is a frequently used dimension reduction method that performs regularization and feature selection (Tibshirani, 1996). Lasso penalizes coefficients of the regression feature, such that coefficients of certain features are shrunk to zero, while coefficients of other features remain non-zero and are considered as selected features.

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{||y - \beta X||_2^2}_{\text{Loss}} + \lambda \underbrace{||\beta||_1}_{\text{Penalty}}$$

The cross-validated Lasso feature selection and model fitting algorithm was performed by the function ”cv.glmnet” implemented in the R packages “glmnet” (Friedman, Hastie, & Tibshirani, 2010), which is a k-fold CV extension of the Lasso function “glmnet”. The following parameter settings were specified: family = "cox", fold = “10” (10-fold CV) and weights equal to the calculated IPW weights. The tuning parameter “lambda” was chosen to match the desired number of features as closely as possible. Therefore, lambda was chosen such that the absolute difference between the target number of features *K* and the number of non-zero features (“nzero”) is minimal.

3.3.6 Study population

Subjects in the lung cancer NCC study were participants of the Beta-Carotene and Retinol Efficacy Trial (CARET) (Sakoda et al., 2012). CARET was a randomized, double-blind, placebo-controlled trial that assessed cancer prevention efficacy and safety of beta-carotene and retinyl palmitate in a high-risk for lung cancer population (Goodman et al., 2004; Omenn, 1997; Omenn et al., 1996). This high-risk population of 18,314 participants is comprised of 7,965 men

and 6,289 women with a heavy smoking history (≥ 20 cigarette pack-years at the time of enrollment), and 4,060 men with extensive occupational asbestos exposure. Whole blood and other specimens were collected from participants at or shortly after enrollment into the trial. In the lung cancer NCC, 316 cases and 316 matched controls were selected. Controls were chosen to match cases in a 1:1 ratio based on age at blood draw (± 4 years), race, sex, exposure type (heavy smoker or asbestos exposure), smoking status at blood draw, and enrollment year (± 2).

3.3.7 Methylation array, quality control, preprocessing, and normalization of methylation data

Whole blood DNA methylation (DNAm) was assessed by the Illumina HumanMethylation850 BeadArray platform at the University of Southern California Epigenomics Core Facility, following standardized protocols described by the manufacturer (Illumina, Inc). For a single DNAm site or cytosine-phosphate-guanine (CpG), DNAm is quantified via the β-value, $\beta = \frac{M}{M+U}$, where *M* and *U* are the methylated and unmethylated signal intensities, respectively. The signal intensities M and U are typically assumed to be gamma-distributed random variables with equal scale parameter (Saadati & Benner, 2014) and it can be shown that their ratio ($\beta = \frac{M}{M+U}$) follows a beta-distribution (see **Appendix VI - Ratio of gamma distributed signals follows beta distribution**). The β-value indicates the methylation extent for a specific CpG and ranges from 0 to 1. Under ideal conditions, a β-value of zero signifies that all alleles in all cells of a sample were unmethylated at that CpG site, while a β-value of one indicates methylation throughout all alleles in all cells at that CpG site (Du et al., 2010).

For the obtained DNAm data, quality control, preprocessing, and normalization was carried out using Bioconductor packages *minfi* and *wateRmelon* (Aryee et al., 2014; Pidsley et

al., 2013). A combination of Noob+β-mixture quantile (BMIQ) was used to perform the within-array normalization, as it has been shown to improve signal sensitivity and outperform other approaches (J. Liu & Siegmund, 2016). In a first step, background fluorescence and dye biases were corrected within an array using *minfi*'s function "preprocessNoob" (Aryee et al., 2014). Next, poor quality samples and probes were excluded. Probes with a median detection p-value $>$ 0.05 and samples with more than 20% of probes that had detection p-values $> 1x10^{-5}$ or irregularities in their distribution of control probes were removed (Wilhelm-Benartzi et al., 2013). Following probe and sample quality control, "BMIQ" function in Bioconductor package *wateRmelon* was implemented to adjust beta-values of type II probes to match the statistical distribution characteristic of type I probes (Teschendorff et al., 2013). BMIQ fits a three-class β-mixture model (unmethylated, hemimethylated, methylated) for each sample on type I and II probes separately, such that each probe is assigned to the class with the highest posterior predictive probability. Each class of type II probe β-values is then normalized to match classes of type II probe. Finally, associations of principal component analysis (PCA) and technical aspects of the array (e.g. batch-effect) were examined, and it was concluded that no further corrections were required (Grieshober et al., 2018).

3.3.8 Simulation study

To demonstrate the applicability of SPOFS in a controlled setting and to benchmark it against Lasso, we considered a series of simulation studies. In our simulation studies, we aimed to mimic the lung cancer NCC study in that we retained its clinical features (matching features) and time-to-event or time-of-censoring information but simulated epigenetic data under different scenarios with varying correlation, effect size and dimensionality. The epigenetic data of the lung cancer NCC study is DNAm, consisting of individual- and CpG-specific β-values. As prior

research has demonstrated that neighboring CpG can exhibit a strong correlation (W. W. Zhang et al., 2015), we desired to generate epigenetic data that exhibits correlation. For the purpose of simplicity, epigenetic data were simulated by drawing observations from a multivariate normal (MVN) distribution. The R packages *MASS* provides the function "mvrnorm" that allows to generate correlated observations by providing a mean vector $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots\}$ and a covariance matrix $\boldsymbol{\Sigma}$. While the mean vector $\boldsymbol{\mu}_{controls}$ for controls is defined as a vector of 0's, the vector of means for cases is defined as a vector of 0's where a defined number of CpGs have a non-zero mean of Δ. The covariance matrix $\boldsymbol{\Sigma}$ for both, cases and controls, is here defined to have variances in the diagonal, covariances with correlation ρ between CpGs that have non-zero means in the cases, and zeros otherwise:

$$D = MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu}_{controls} = \{0,0,0,0,0,0 \ldots\}$$

$$\boldsymbol{\mu}_{cases} = \{0, \Delta, 0, \Delta, 0,0 \ldots\}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 & \\ 0 & \sigma_2^2 & 0 & \rho\sigma_2\sigma_4 & 0 & \\ 0 & 0 & \sigma_3^2 & 0 & 0 & \ldots \\ 0 & \rho\sigma_4\sigma_2 & 0 & \sigma_4^2 & 0 & \\ 0 & 0 & 0 & 0 & \sigma_5^2 & \\ & & \ldots & & & \ddots \end{bmatrix}$$

$$\sigma^2 \sim Unif(0,10)$$

While reusing clinical data from the NCC in CARET, epigenetic data sets were created for different combinations of the number of simulated CpGs, number of CpGs with non-zero means for cases, magnitude of non-zero means Δ, and correlation ρ summarized in the following scenarios:

Table 4 Summary describing parameter setting for simulation scenarios assessing the influence of correlation.

Parameter \ Scenario	Number of simulated CpGs (P)	Number of CpGs with non-zero means for cases	Magnitude of non-zero means (Δ)	Correlation (ρ)
Scenario 1	1000	5	0.3	0.9
Scenario 2	1000	5	0.3	0.8
Scenario 3	1000	5	0.3	0.7
Scenario 4	1000	5	0.3	0.6
Scenario 5	1000	5	0.3	0.5
Scenario 6	1000	5	0.3	0

Table 5 Summary describing parameter setting for simulation scenarios assessing the influence of magnitude of introduce differences.

Parameter \ Scenario	Number of simulated CpGs (P)	Number of CpGs with non-zero means for cases	Magnitude of non-zero means (Δ)	Correlation (ρ)
Scenario 1	1000	5	0.3	0.9
Scenario 7	1000	5	0.5	0.9
Scenario 8	1000	5	0.7	0.9

Table 6 Summary describing parameter setting for simulation scenarios assessing the influence of the number of CpGs simulated with non-zero means for cases.

Parameter \ Scenario	Number of simulated CpGs (P)	Number of CpGs with non-zero means for cases	Magnitude of non-zero means (Δ)	Correlation (ρ)
Scenario 1	1000	5	0.3	0.9
Scenario 9	50	5	0.3	0.9
Scenario 10	10000	5	0.3	0.9

3.4 Results

With the goal in mind of being able to validate results, we divided data from the lung cancer NCC study and each simulated data set into two parts. Specifically, 75% of the data were designated for model building and performance assessment (training-and-validation set), while 25% of the data were held back (testing set) to evaluate final models. However, results can be influenced by the way the data were split, especially within the context of a time-to-event analysis. To compensate for any potential effects resulting from splitting the data, all results are generated by performing a random split into training-and-validation set and testing set for different seed settings. As data of such high dimensionality is susceptible to model over-fitting, models within each training-and-validation set were fit by performing 10-fold cross-validation (10-fold CV) to reduce potential over-fitting effects and to generate models that are generalizable.

3.4.1 NCC study in CARET

With the objective of identifying a set of CpGs that collectively predict lung cancer in the high-risk population comprising the lung cancer NCC study, we applied both methods (SPOFS and Lasso) to training-and-validation sets and evaluated model performance of each best model in the testing set. Models including 5, 15, and 30 epigenetic features (CpGs) were developed for ten different seeds ($\text{seed} = \{1,2,\dots,10\}$) to attenuate the potential effects of splitting data sets into training-and-validation and testing sets.

The initial step of SPOFS was performed to rank-order $> 850{,}000$ CpGs on the basis of their c-index. For the purpose of reducing dimensionality, the $1{,}000$ CpGs with the highest c-indices were retained and considered for subsequent steps. Based on intermediate results of a variety of different weighting functions (see **Appendix V - List of potential weighting functions**) used in step 7) of SPOFS, we decided to apply a sigmoid function for calculating weights used for optimizing selection probabilities, defined as:

$$f\left(\Delta_k^{(m)}\right) = \frac{1}{1 + \exp(-5\Delta_k^{(m)})}$$

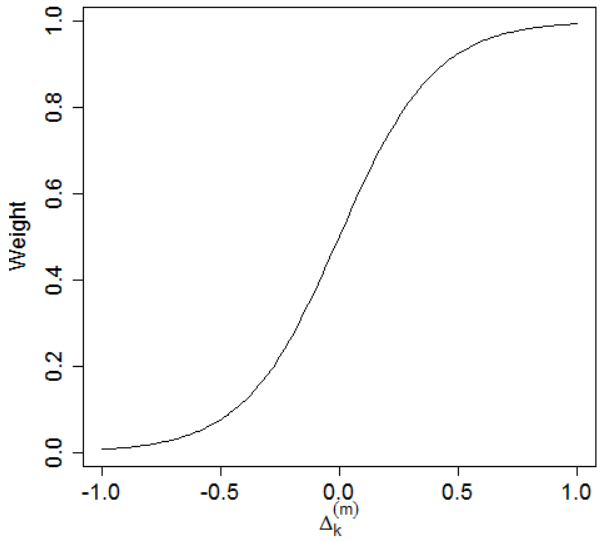


Figure 7 Weighting function defined as Sigmoid function.

For the purpose of comparison, the 10-fold cross-validated Lasso function "cv.glmnet" was executed with default settings and the penalty coefficient lambda was chosen such that the absolute difference between non-zero coefficients and the user-defined number of features (5, 15, 30 CpGs) was minimized. The results of both methods, SPOFS and Lasso, are provided in Figure 8A-C. In addition, results were also obtained by choosing the penalty coefficient lambda as "lambda.min", where Lasso's mean cross-validated error is minimized, and the number of non-zero feature is automatically chosen (Figure 8D).

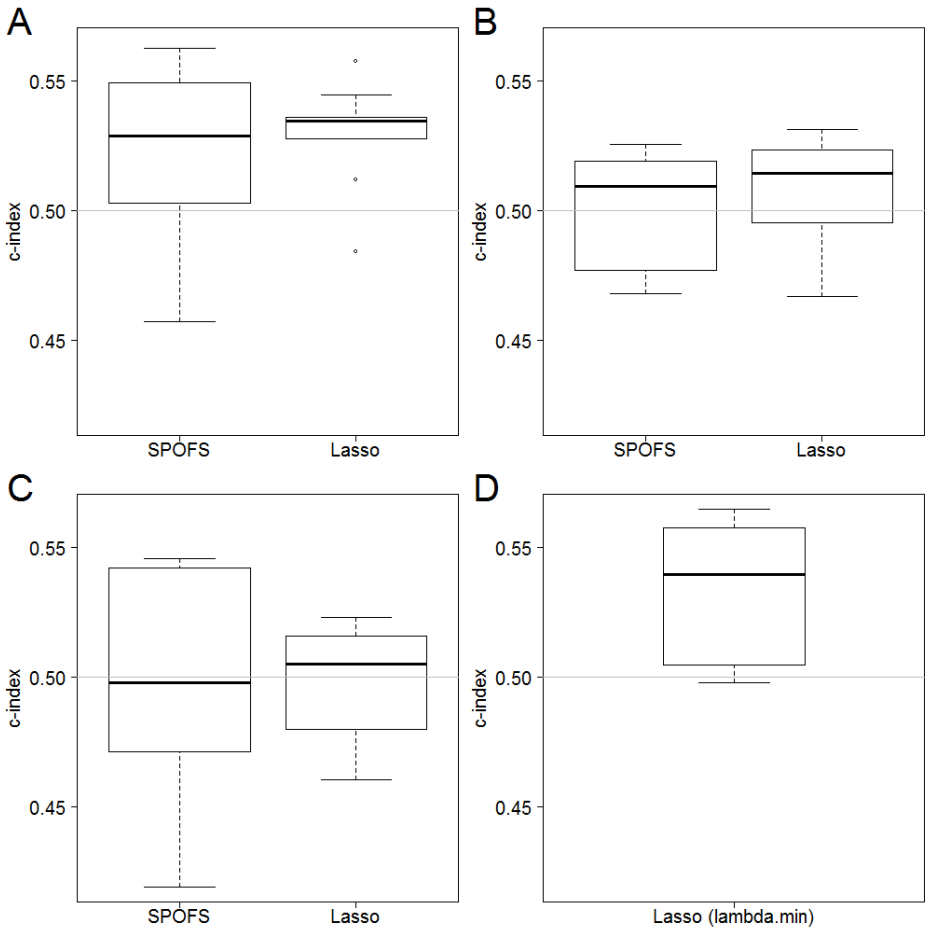


Figure 8 Prediction performance of SPOFS and Lasso in lung cancer NCC study.

Subfigure A-C illustrate the prediction performance of models including 5, 15 and 30 CpGs, respectively, that were identified by SPOFS and Lasso. Subfigure D demonstrates the prediction performance of Lasso when the penalty coefficient lambda is chosen as "lambda.min".

It can be observed that both methods, SPOFS and Lasso, achieved similar c-index, which were only marginally better than 0.50 (see Figure 8A-C) in the testing data set when including 5, 15, and 30 CpGs. Further, it can be observed that the average c-index decreases, as the number of CpGs in the model increases; potentially a result of overfitting. In Figure 8D, prediction performance of Lasso is visualized when the penalty coefficient lambda is chosen "lambda.min". This allows Lasso to automatically determine the number of features based on the minimized mean cross-validated error. Here, the number of features automatically selected by Lasso is equal

to zero (7 out of 10 times) and one (3 out of 10 times). This reveals that Lasso's best prediction performance is achieved here, when either no or one CpG is included in a model that controls for the matching variables.

These results suggest that either critically important assumptions were violated or that blood-derived DNAm data do not provide a substantial amount of information to predict lung cancer. To further investigate if the results shown above are due to a lack of signal in the data or due the inapplicability of both methods to the lung cancer NCC data, we conducted a series of simulation studies under a variety of different simulation scenarios.

3.4.2 Simulated data

With the goal of demonstrating that our proposed method SPOFS can competitively identify predictive features, we evaluated its performance based on simulated epigenetic data under a variety of scenarios (see Table 4, Table 5, Table 6). As previously, 75% of the data are designated for model building and performance assessment (training-and-validation set), while 25% of the data were held back (testing set) to evaluate final models. For each simulation scenario, data were divided into both sets after specifying 100 different seeds. We then applied SPOFS and Lasso to the training-and-validation sets and evaluated the performance of selected features in the held-back testing data sets. We also evaluated the "Upper limit" of performance, which indicates the performance of models including all CpGs that were simulated to have an effect on time-to-diagnosis. The "Upper limit" serves as a point of reference and represents the scenario that can be thought of as the "gold-standard" for prediction performance.

Figure 9 visualizes the prediction performance of models identified by SPOFS and Lasso in scenario 1, where data was simulated with a combination of strong correlation ($\rho = 0.9$) and weak signal ($\Delta = 0.3$) of epigenetic features. Here, we demonstrate that SPOFS achieves a

significantly better prediction performance based on its selected features compared to Lasso (one-sided Wilcoxon rank-sum test p-value $= 0.00028$).

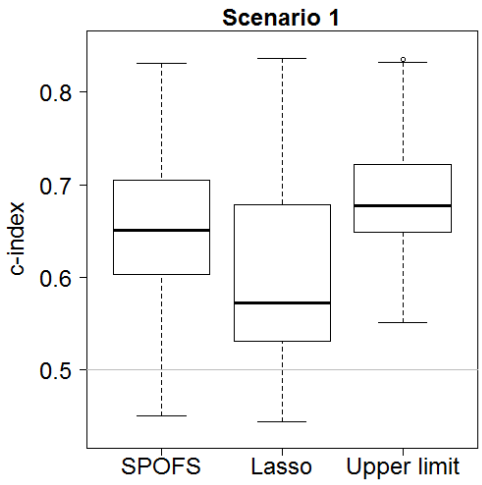


Figure 9 Prediction performance of features selected by SPOFS and Lasso in scenario 1.

In scenarios 1-6, we examined the influence of the simulated correlation on the ability of SPOFS and Lasso to identify an optimal set of features. Therefore, we reduced the correlation stepwise from 0.9 to 0.5 and as well as setting the correlation equal to zero. The results are shown in Figure 10, illustrating the prediction performance of features selected by SPOFS and Lasso, and each scenario's "Upper limit". It can be observed, that in scenario 6, where no correlation was simulated, SPOFS and Lasso demonstrate comparable prediction results. However, it can be observed that with increasing correlation (scenario 5-1) SPOFS performance to predict lung cancer exceeds Lasso's. The difference in prediction performance is significant at the correlation of $\rho = 0.8$ (scenario 2, one-sided Wilcoxon rank-sum test p-value$= 0.00916$), while the difference in prediction performance is not significant at the correlation of $\rho = 0.7$ (scenario 3, one-sided Wilcoxon rank-sum test p-value$= 0.16632$).

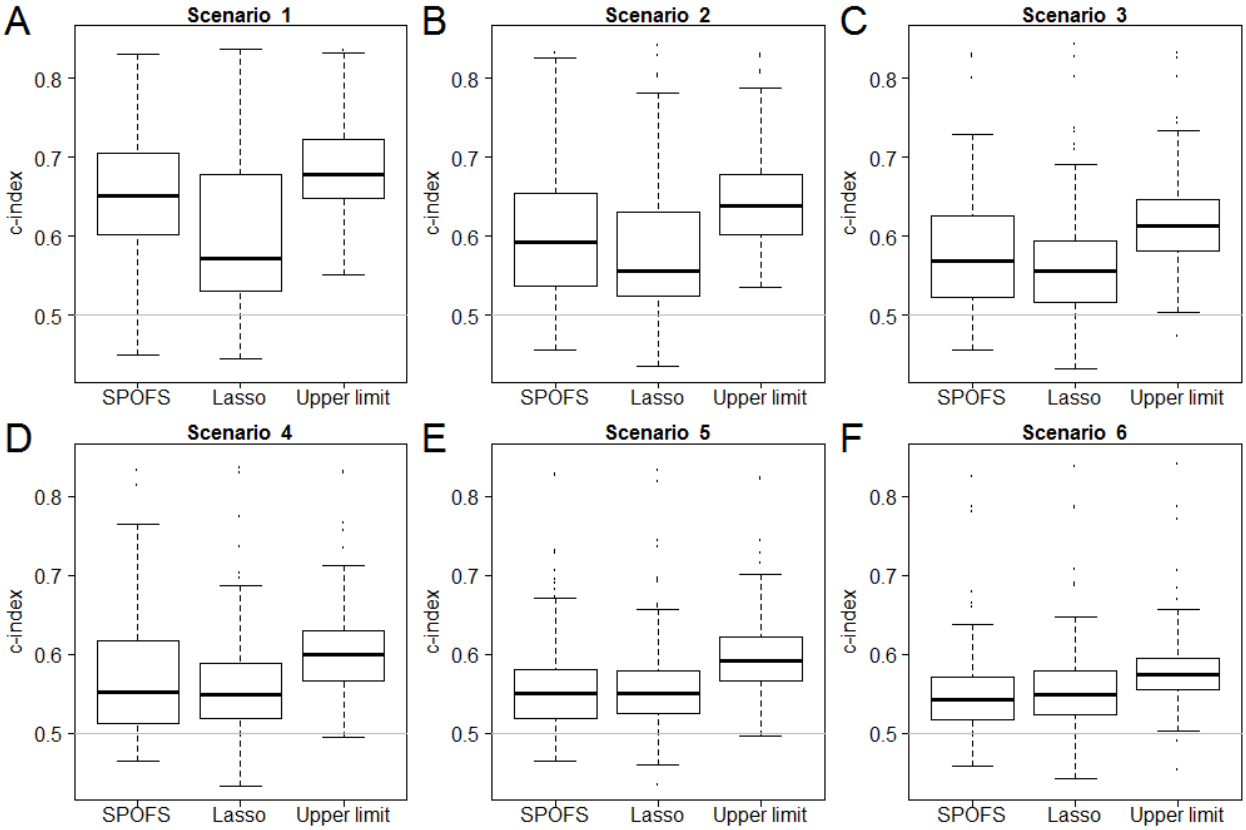


Figure 10 Prediction performance of features selected by SPOFS and Lasso.

A) scenario 1 with $\rho = 0.9$; B) scenario 2 with $\rho = 0.8$; C) scenario 3 with $\rho = 0.7$; D) scenario 4 with $\rho = 0.6$; E) scenario 5 with $\rho = 0.5$; F) scenario 6 with $\rho = 0$.

Next, we investigated the influence of the magnitude of the introduced differences on the ability of SPOFS and Lasso to identify an optimal set of features. Therefore, we generated simulated data sets with $\Delta = \{0.5, 0.7\}$ (scenario 7 and 8). The results are shown below (Figure 11), illustrating prediction performances of feature selected by SPOFS and Lasso, and each scenario's "Upper limit". It can be observed that with increasing simulated differences between cases and controls, the ability to identify an optimal set of CpGs increases for both methods, SPOFS and Lasso, and prediction performances are more similar to the "Upper limit". While the differences between prediction performances of SPOFS and Lasso are marginally significant in

scenario 7 (one-sided Wilcoxon rank-sum test p-value= 0.05462), there is no significant difference observed in scenario 8 (one-sided Wilcoxon rank-sum test p-value= 0.29845).

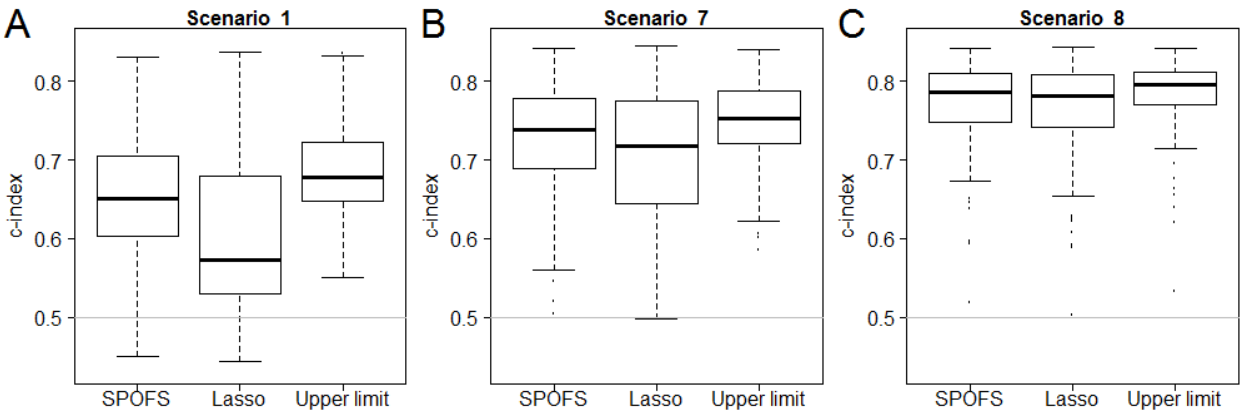


Figure 11 Prediction performance of features selected by SPOFS and Lasso.

A) scenario 1 with $\Delta = 0.3$; B) scenario 7 with $\Delta = 0.5$; C) scenario 8 with $\Delta = 0.7$.

Finally, we explored the influence of the total number of CpGs simulated on the ability of SPOFS and Lasso to identify an optimal set of features. Therefore, for scenario 9 and 10 we generated data sets by simulating 50 and 10,000 CpGs, respectively. It can be observed in Figure 12 that an increasing number of features introduces an increased challenge to identify an optimal subset of CpGs to predict lung cancer. For neither, a reduced number of CpGs nor an increased number of CpGs, the difference in prediction performance for SPOFS and Lasso was significant (scenario 9, one-sided Wilcoxon rank-sum test p-value= 0.11301; scenario 10, one-sided Wilcoxon rank-sum test p-value= 0.60642)

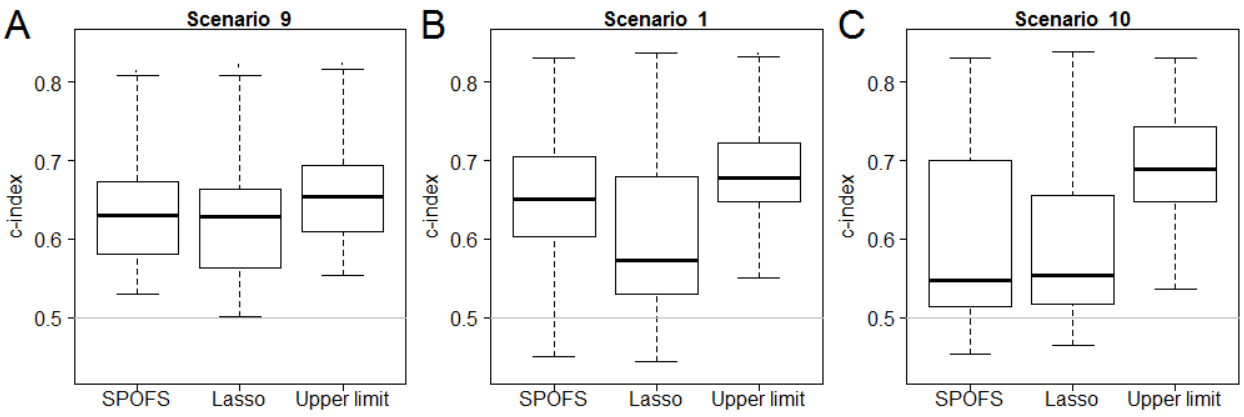


Figure 12 Prediction performance of features selected by SPOFS and Lasso.

A) scenario 9 with a total number of 50 CpGs simulated; B) scenario 1 with a total number of 1000 CpGs simulated; C) scenario 10 with a total number of 10.000 CpGs simulated.

3.5 Discussion

Motivated by goal of feature selection in the context of high dimensional epigenetic data collected on subjects in a NCC study of lung cancer risk, we present a heuristic algorithm, where feature's selection probabilities are iteratively optimized, while monitoring overall performance to identify an optimal subset of features. We employ the frequently used method Lasso to benchmark our results. In the lung cancer NCC study neither SPOFS nor Lasso provided satisfying feature selection results, however in simulated data, we demonstrate that SPOFS is a competitive method compared to Lasso. While both algorithms achieve highly comparable results, we have established that SPOFS outperforms Lasso in scenarios involving a combination of strong correlation and weak signal.

There exist several potential reasons that can explain the lack of success in the lung cancer NCC study. One possible reason for the modest performance of both methods is the absence of predictive signal in the data. As neither method is able to identify a set of CpGs that improves the prediction performance of lung cancer in this lung cancer NCC study, it can be

speculated that the data do not contain any predictive CpGs for lung cancer. This lack of predictive signal can be the consequence of the time differences from blood-draw to lung cancer diagnosis. The median time from blood-draw to lung cancer diagnosis was 4.4 years with the time at risk (time from blood-draw to lung cancer diagnosis) ranging up to more than 10 years. It can be questionable to expect the identification of reliable biomarkers several years prior to the lung cancer diagnosis. To achieve reliable predictions, we might require epigenetic data from individuals who had their blood drawn more proximal to their diagnosis of lung cancer. Further, the lack of predictive signal could be a result of an averaging effect of histology types. It has been shown that the histology types of lung cancer exhibit distinct DNA methylation signatures (Grieshober et al., 2018). Because the objective was to identify a pre-diagnostic biomarker for lung cancer, we did not control for histology types of lung cancer, as they involve information that is not available prior to lung cancer diagnosis. However, we potentially fail to observe histology type specific pattern that are predictive of lung cancer, by averaging over them, as we do not incorporate them in the model building process. In future work, these speculations should be investigated by controlling for the histology type.

Another possible explanation for lack of success in identifying predictive CpGs in the lung cancer NCC study involves the choice of model. Here, both methods were carried out in the framework of Cox proportional hazards model, due to its popularity, robustness (Sestelo, 2017) and ability to incorporate IPW calculated from the cohort. However, there exist different models that may be more suitable for the purposes of this study. If we assume that an increased hazard can be expected to be observed for individuals whose blood was taken more proximal to the diagnosis of lung cancer, then appealing alternatives include accelerated failure time models (AFT model), such as Weibull models. Lastly, the modest prediction performance of both

methods, SPOFS and Lasso, can potentially be the result of the metric used for model evaluation. Here, due to its popularity, we evaluated prediction performance of models based on their concordance index, but we acknowledge that other common metrics, such as Brier score and mean absolute error, to assess prediction performance are potentially more suitable.

Nevertheless, in the simulated data, we have demonstrated that SPOFS is a competitive method compared to Lasso. While both algorithms achieve comparable results, we have established that Lasso's performance can be compromised in scenarios involving a combination of weak signal and strong correlation in the data. This coincides with previous literature, which has reported that Lasso requires weak correlation and strong signal conditions on the design matrix to achieve high selection accuracy (Y. L. Zhang, 2017). This explains the performance of Lasso in simulated scenarios, where a combination of weak signal and strong correlation was specified for the generation of data. However, prior research has demonstrated that neighboring CpGs can exhibit a strong correlation (W. W. Zhang et al., 2015) and therefore, Lasso may be less applicable in such scenarios. In future iterations it is advisable to compare prediction performance of SPOFS under a variety of scenarios to more advanced and sophisticated methods, such as adaptive Lasso or elastic net (Zou, 2006; Zou & Hastie, 2005).

The simulated data in this manuscript were generated from a Multivariate Normal (MVN) distribution for the purpose of simplicity, as the function "mvrnorm" of the R package *MASS* facilitates the option to provide covariance matrix. However, for a single CpG (epigenetic feature), DNAm is commonly quantified via β-values, which follow a beta distribution. To our knowledge, there exists no available R packages that assists one in generating correlated observations from beta distributions. However, it can be shown (see **Appendix VII - Generation of correlated beta values**) that correlated observations from a beta distribution

($Y_1 \sim beta(c_1, c_2), Y_2 \sim beta(c_3, c_4)$) can be obtained by taking the sum of two marginal Dirichlet distributed observations, where the concentration parameters of the Dirichlet distribution are functions of the target correlation and the shape parameters $c_{\{1,2,3,4\}}$ of the beta distributions. Therefore, one is required to specify the shape parameter $c_{\{1,2,3\}}$ of the beta distribution such that $c_3 < c_1 + c_2$ (shape parameter c_4 will be determined based on the specified correlation). Next, the concentration parameters of the Dirichlet distribution can be calculated and used to generate observations from a Dirichlet distribution $\boldsymbol{X} \sim dir(\alpha_1, \alpha_2, \alpha_j, \alpha_0)$. Finally, the sums of two marginal Dirichlet distributed observations ($Y_1 = X_1 + X_j$ and $Y_1 = X_2 + X_j$) allow one to obtain beta-distributed observations that exhibit a correlation ρ. This procedure can be scaled up to facilitate a correlation structure of multiple CpGs. Detailed proofs, calculations and a sampling procedure (including an example) can be found in **Appendix VII - Generation of correlated beta values**.

We have demonstrated the comparability of SPOFS' and Lasso's feature selection performance, but it is important to point out that the cross-validated feature selection performed by Lasso provides the opportunity to select the tuning parameter λ such that the mean cross-validated error is minimized ("lambda.min") and the number of non-zero feature is automatically chosen. In contrast, the number of epigenetic features is currently a necessary input for SPOFS. However, this presents an opportunity for future enhancements of this method, in which a range of number of features is scanned and evaluated.

Further, it deserves mentioning that while SPOFS is designed to optimize a set of features for prediction, it does not evaluate all possible feature combinations (i.e., $\binom{P}{K}$), and therefore does not ensure a globally optimal set of features. The computational burden associated with the evaluation of all possible feature combinations is unmanageable and prohibitive. To reduce the

computational burden and run time of SPOFS, we utilize an initial filtering step to reduce dimensionality and additionally implemented a parallel processing of the cross-validated steps (steps 3-5). However, the initial filtering step involves a risk of potentially excluding features that are predictive in the combination with other features. An additional limitation that can also be considered as an advantage is the fact that SPOFS does not have a predetermined completion, meaning that the user is required to define a termination condition. But this also means that there always exists a chance of identifying a better performing set of epigenetic features with additional run time.

Furthermore, SPOFS identifies a list of well-performing sets of features. This list provides the opportunity of combining multiple models in a model ensemble in order to leverage predictions of each individual model. Previous literature has demonstrated that the combination of multiple models in an ensemble outperforms the best participating single model in most scenarios (Weigel, Liniger, & Appenzeller, 2008).

While SPOFS was developed here for a time-to-event analysis in the context of a nested case-control study, its methodology is applicable to most circumstances where a selection of features evaluated on some prediction metric is desired. In future iterations of SPOFS there exist several enhancements that can be incorporated to potentially improve prediction performance. In this manuscript we only have considered main effect combinations of clinical and epigenetic features. SPOFS' methodology can be extended by incorporating interactions between and within clinical and epigenetic features. Another possible extension of SPOFS includes the combination of different evaluation metrics, as previously demonstrated in IDOL (Koestler et al., 2016), where a composite measure was generated by combining of R^2 and root mean square errors ($RMSE$).

While assessment of different simulation scenarios provides promising results, the evaluation of SPOFS in a real data set remains to be demonstrated, as data from the lung cancer NCC study did not provide satisfying results for either method. This also provides the opportunity for incorporating additional (high-dimensional) omic data types, such as gene expression.

3.6 Conclusion

Motivated by the desire to identify a pre-diagnostic biomarker consisting of a subset of epigenetic features (CpGs) that collectively predict lung cancer in the high-risk population of a nested case-control study in CARET, we here propose a heuristic feature selection method SPOFS. While neither SPOFS nor Lasso were able to identify a promising biomarker in the lung cancer nested case-control study of CARET, we have demonstrated in simulated data that SPOFS achieves competitive results compared to Lasso and exceeds the performance of lasso in certain scenarios involving a combination of weak signal and strong correlation structure.

3.7 Contribution

In this project I devised and implemented the SPOFS algorithm, preprocessed data from the lung cancer NCC study, generated simulated epigenetic data, estimated weights for IPW, analyzed data sets and wrote the manuscript.

4 Summary and Future Directions

The discovery and establishment of predictive markers is a crucial part of the process for the early detection and prognosis of many diseases. Being able to detect a life-threatening disease such as cancer at an early stage can save lives and drastically improve quality of life among patients. Treatments have been shown to be more likely to be effective and efficient at an early stage of the disease progression and ensure a higher potential of cure (WHO, 2007). Patients' responses to treatments can vary by individuals and employment of predictive markers for disease prognosis grant the gateway to personalized medicine, where treatments are specific to an individual patient based on their predicted response or risk of a disease (FORUM - Academy of Medical Sciences, 2015).

Motivated by the potential to improve risk-prediction for prostate cancer, the Prostate Cancer Dream Challenge was launched 2015 with the objective to develop prognostic models for predicting overall survival. We have demonstrated that our curated, ad-hoc, feature selection (CAFS) strategy is able to identify clinically important risk-predictors of patient survival. Several of the identified predictors were new features created by strategically merging collections of weak predictors. The predictive features identified by CAFS were utilized to our ensemble-based Cox proportional hazards regression framework and we established prediction models that outperformed the gold-standard prediction model for prostate cancer survival (at that time) and ranked among the top models developed by teams from around the world.

A considerable limitation of the Prostate Cancer Dream Challenge is the utilization of clinical features exclusively, as no omic or molecular data were provided for this competition. The incorporation of omic features in the prediction model development provides a promising opportunity to further improve prediction performances of patient's prognosis. Such epigenetic

marks (i.e., DNA methylation) are generally investigated in epigenome-wide association studies (EWAS) to identify biomarkers that are associated with some exposure(s) or phenotype(s). As for any study, the identification of these biomarkers depends on an adequate sample size and associated power. However, due to the complex nature of DNA methylation (DNAm), there is a lack of tools and methods for power evaluation of EWAS. With the goal to fill this gap, we have developed pwrEWAS, a user-friendly tool for comprehensive power estimation for EWAS. With pwrEWAS we provide a publicly available R package including a user-friendly point-and-click interface for the power evaluation of EWAS. The package facilitates power estimation for two-group comparisons of DNAm by applying a semi-parametric simulation-based approach. With pwrEWAS, we provide a user-friendly tool that assists researchers in the design and planning of EWAS.

We then revert back to the development of prediction models and discuss the integration of these epigenetic biomarkers to enhance prediction performance by clinical features. Motivated by the challenge of feature selection in the context of high dimensional data, we introduced a flexible Selection Probability Optimization method for Feature Selection (SPOFS). SPOFS is designed to identify an optimal subset of epigenetic features and combine these potentially weak predictors in an epigenetic biomarker that improves the prediction performance of prognostic models with clinical features. We have demonstrated that SPOFS is a competitive feature selection method compared to Lasso, while outperforming it in scenarios involving a combination of weak signal and strong correlation. As considerable correlations can be expected in most omic data sets, we hypothesize that SPOFS is more suitable in such scenarios.

In future work, we want to extend and improve each algorithm and tool. With the ongoing development of statistical methods and models, we want to explore the prediction

performance of CAFS in the context of models besides the Cox proportional-hazards model. Additionally, we desire to address the bias of CAFS in its current implementation to including interaction terms over main effect. This could be achieved by penalizing the inclusion of interaction terms. The current implementation of pwrEWAS allows for a variety of extension, including comparison of multiple groups. We also would like to incorporate the option for researchers to upload different tissue types that are specific to their study and address the potential change of CpG dispersion as a result of some phenotype or exposure. Further, we wish to implement additional methods for differential methylation analysis, as methods improve and develop. Lastly, we want to optimize the set of features selected by SPOFS by incorporating interactions between and within clinical and epigenetic features and automatize the choice of the number of features. Additionally, we would like to combine multiple evaluation metrics with the goal to improve and stabilize the prediction performance. Finally, we want to evaluate SPOFS' performance in the context of different models and compare the performance of SPOFS with more advanced and sophisticated methods.

Ultimately, we would like to apply a combination of both feature selection methods, CAFS and SPOFS. For a given data set that provides a great number of clinical features and one or more types of omic data, we desire to apply CAFS to the clinical data to identify clinically important risk-predictors, while SPOFS can be applied to one or a combination of different omic data sets (e.g. DNA methylation and gene expression) to identify optimal set(s) of omic features that improve the ability to predict some outcome or response.

References

Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Le Calvez-Kelm, F., Kaaks, R., . . . Herceg, Z. (2016). Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics, 8*(5), 599-618. doi:10.2217/epi-2016-0001

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics, 30*(10), 1363-1369. doi:10.1093/bioinformatics/btu049

Barfield, R. T., Kilaru, V., Smith, A. K., & Conneely, K. N. (2012). CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics, 28*(9), 1280-1281. doi:10.1093/bioinformatics/bts124

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L. X., Shen, R., & Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium (R) assay. *Epigenomics, 1*(1), 177-200. doi:10.2217/Epi.09.14

Biesheuvel, C. J., Vergouwe, Y., Oudega, R., Hoes, A. W., Grobbee, D. E., & Moons, K. G. (2008). Advantages of the nested case-control design in diagnostic research. *Bmc Medical Research Methodology, 8*. doi:Artn 48 10.1186/1471-2288-8-48

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther, 69*(3), 89-95. doi:10.1067/mcp.2001.113989

Blaise, B. J., Correia, G., Tin, A., Young, J. H., Vergnaud, A. C., Lewis, M., . . . Ebbels, T. M. (2016). Power Analysis and Sample Size Determination in Metabolic Phenotyping. *Anal Chem, 88*(10), 5179-5188. doi:10.1021/acs.analchem.6b00188

Bramsen, J. B., Rasmussen, M. H., Ongen, H., Mattesen, T. B., Orntoft, M. B. W., Arnadottir, S. S., . . . Andersen, C. L. (2017). Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell Reports, 19*(6), 1268-1280. doi:10.1016/j.celrep.2017.04.045

Cai, T., & Zheng, Y. (2012). Evaluating prognostic accuracy of biomarkers in nested case-control studies. *Biostatistics, 13*(1), 89-100. doi:10.1093/biostatistics/kxr021

Chalise, P. a., Dai, J. a., Ellis, S. a., Fridley, B. a., Graw, S. a., Koestler, D. a., . . . Usset, J. (2016). JayhawksProstateDream: First release (PCDC submission). doi:10.5281/zenodo.49063

Chang, K., Kong, Y. Y., Dai, B., Ye, D. W., Qu, Y. Y., Wang, Y., . . . Li, G. X. (2015). Combination of circulating tumor cell enumeration and tumor marker detection in predicting prognosis and treatment effect in metastatic castration-resistant prostate cancer. *Oncotarget, 6*(39), 41825-41836. doi:10.18632/oncotarget.6167

Chen, H., Li, G., Chan, Y. L., Chapman, D. G., Sukjamnong, S., Nguyen, T., . . . Oliver, B. G. (2018). Maternal E-Cigarette Exposure in Mice Alters DNA Methylation and Lung Cytokine Expression in Offspring. *American Journal of Respiratory Cell and Molecular Biology, 58*(3), 366-377. doi:10.1165/rcmb.2017-0206RC

Chen, K. N. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 63*, 791-809. doi:Doi 10.1111/1467-9868.00313

Ching, T., Huang, S. J., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *Rna, 20*(11), 1684-1696. doi:10.1261/rna.046011.114

Choi, J., Park, S., Yoon, Y., & Ahn, J. (2017). Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. *Bioinformatics, 33*(22), 3619-3626. doi:10.1093/bioinformatics/btx487

Du, P., Zhang, X. A., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L. F., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *Bmc Bioinformatics, 11*. doi:Artn 587 10.1186/1471-2105-11-587

Feng, S., Wang, S. C., Chen, C. C., & Lan, L. (2011). GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *Bmc Genetics, 12*. doi:Artn 12 10.1186/1471-2156-12-12

Fizazi, K., Higano, C. S., Nelson, J. B., Gleave, M., Miller, K., Morris, T., . . . Moul, J. W. (2013). Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology, 31*(14), 1740-1747. doi:10.1200/JCO.2012.46.4149

FORUM - Academy of Medical Sciences. (2015). Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1-22.

Goodman, G. E., Thornquist, M. D., Balmes, J., Cullen, M. R., Meyskens, F. L., Omenn, G. S., . . . Williams, J. H. (2004). The beta-carotene and retinol efficacy trial: Incidence of lung cancer and cardiovascular disease mortality during 6-year follow-up after stopping beta-carotene and retinol supplements. *Journal of the National Cancer Institute, 96*(23), 1743-1750. doi:10.1093/jnci/djh320

Grace, H. H., & Li, Y. (2017). Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review. *Appl Math, 32*(4), 379-396.

Graw, S., Henn, R., Thompson, J. A., & Koestler, D. C. (2019). pwrEWAS: a user-friendly tool for comprehensive power estimation for epigenome wide association studies (EWAS). *Bmc Bioinformatics, 20*(1), 218. doi:10.1186/s12859-019-2804-7

Grieshober, L., Graw, S., Barnett, M. J., Thornquist, M. D., Goodman, G. E., Chen, C., . . . Doherty, J. A. (2018). Methylation-derived Neutrophil-to-Lymphocyte Ratio and Lung Cancer Risk in Heavy Smokers. *Cancer Prev Res (Phila), 11*(11), 727-734. doi:10.1158/1940-6207.CAPR-18-0111

Guinney, J., Wang, T., Laajala, T. D., Winner, K. K., Bare, J. C., Neto, E. C., . . . Comm, P. C. C. D. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. *Lancet Oncology, 18*(1), 132-142. doi:10.1016/S1470-2045(16)30560-5

Guo, Y., Zhao, S., Li, C. I., Sheng, Q., & Shyr, Y. (2014). RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. *Cancer Inform, 13*(Suppl 6), 1-5. doi:10.4137/CIN.S17688

Haibe-Kains, B., Desmedt, C., Sotiriou, C., & Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics, 24*(19), 2200-2208. doi:10.1093/bioinformatics/btn374

Halabi, S., Lin, C. Y., Kelly, W. K., Fizazi, K. S., Moul, J. W., Kaplan, E. B., . . . Small, E. J. (2014). Updated Prognostic Model for Predicting Overall Survival in First-Line Chemotherapy for Patients With Metastatic Castration-Resistant Prostate Cancer. *Journal of Clinical Oncology, 32*(7), 671-+. doi:10.1200/Jco.2013.52.3696

Hall, E., Volkov, P., Dayeh, T., Esguerra, J. L. S., Salo, S., Eliasson, L., . . . Ling, C. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biology, 15*(12). doi:ARTN 522 10.1186/s13059-014-0522-z

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., . . . Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *Plos Genetics, 14*(8). doi:ARTN e1007544 10.1371/journal.pgen.1007544

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., . . . Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat Genet, 43*(8), 768-775. doi:10.1038/ng.865

Hong, S. R., Jung, S. E., Lee, E. H., Shin, K. J., Yang, W. I., & Lee, H. Y. (2017). DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. *Forensic Sci Int Genet, 29*, 118-125. doi:10.1016/j.fsigen.2017.04.006

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schonfels, W., Ahrens, M., . . . Hampe, J. (2014). Obesity accelerates epigenetic aging of human liver. *Proc Natl Acad Sci U S A, 111*(43), 15538-15543. doi:10.1073/pnas.1412759111

Hung, H., & Chiang, C. T. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics-Revue Canadienne De Statistique, 38*(1), 8-26.

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., . . . London, S. J. (2016). Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet, 9*(5), 436-447. doi:10.1161/CIRCGENETICS.116.001506

Kawai, T., Yamada, T., Abe, K., Okamura, K., Kamura, H., Akaishi, R., . . . Hata, K. (2015). Increased epigenetic alterations at the promoters of transcriptional regulators following inadequate maternal gestational weight gain. *Scientific Reports, 5*. doi:ARTN 14224 10.1038/srep14224

Kim, R. S. (2013). Analysis of Nested Case-Control Study Designs: Revisiting the Inverse Probability Weighting Method. *Commun Stat Appl Methods, 20*(6), 455-466. doi:10.5351/CSAM.2013.20.6.455

Koestler, D. C., Jones, M. J., Usset, J., Christensen, B. C., Butler, R. A., Kobor, M. S., . . . Kelsey, K. T. (2016). Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics, 17*, 120. doi:10.1186/s12859-016-0943-7

Kular, L., Liu, Y., Ruhrmann, S., Zheleznyakova, G., Marabita, F., Gomez-Cabrero, D., . . . Jagodic, M. (2018). DNA methylation as a mediator of HLA-DRB1(star)15:01 and a protective variant in multiple sclerosis. *Nature Communications, 9*. doi:ARTN 2397 10.1038/s41467-018-04732-5

Kulis, M., & Esteller, M. (2010). DNA Methylation and Cancer. *Epigenetics and Cancer, Pt A, 70*, 27-56. doi:10.1016/S0065-2660(10)70002-X

Langie, S. A. S., Moisse, M., Declerck, K., Koppen, G., Godderis, L., Vanden Berghe, W., . . . De Boever, P. (2017). Salivary DNA Methylation Profiling: Aspects to Consider for

Biomarker Identification. *Basic & Clinical Pharmacology & Toxicology, 121*, 93-101. doi:10.1111/bcpt.12721

Langie, S. A. S., Moisse, M., Szic, K. S. V., Van der Plas, E., Koppen, G., De Prins, S., . . . De Boever, P. (2018). GLI2 promoter hypermethylation in saliva of children with a respiratory allergy. *Clinical Epigenetics, 10*. doi:ARTN 50 10.1186/s13148-018-0484-1

Lee, H. S., & Krischer, J. P. (2017). A new framework for prediction and variable selection for uncommon events in a large prospective cohort study. *Model Assist Stat Appl, 12*(3), 227-237. doi:10.3233/MAS-170397

Li, D. M., Xie, Z. D., Le Pape, M., & Dye, T. (2015). An evaluation of statistical methods for DNA methylation microarray data analysis. *Bmc Bioinformatics, 16*. doi:ARTN 217 10.1186/s12859-015-0641-x

Liang, S., Ma, A., Yang, S., Wang, Y., & Ma, Q. (2018). A Review of Matched-pairs Feature Selection Methods for Gene Expression Data Analysis. *Comput Struct Biotechnol J, 16*, 88-97. doi:10.1016/j.csbj.2018.02.005

Liddell, F. D. K., Mcdonald, J. C., & Thomas, D. C. (1977). Methods of Cohort Analysis - Appraisal by Application to Asbestos Mining. *Journal of the Royal Statistical Society Series a-Statistics in Society, 140*, 469-491. doi:Doi 10.2307/2345280

Liu, J., & Siegmund, K. D. (2016). An evaluation of processing methods for HumanMethylation450 BeadChip data. *Bmc Genomics, 17*. doi:ARTN 469 10.1186/s12864-016-2819-7

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., . . . Feinberg, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology, 31*(2), 142-147. doi:10.1038/nbt.2487

Logue, M. W., Smith, A. K., Wolf, E. J., Maniates, H., Stone, A., Schichman, S. A., . . . Miller, M. W. (2017). The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics, 9*(11), 1363-1371. doi:10.2217/epi-2017-0078

Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., . . . Tonisson, N. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology, 15*(4). doi:ARTN r54 10.1186/gb-2014-15-4-r54

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of 2 Random Variables Is Stochastically Larger Than the Other. *Annals of Mathematical Statistics, 18*(1), 50-60. doi:DOI 10.1214/aoms/1177730491

Mark, S. D., & Katki, H. A. (2006). Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. *Journal of the American Statistical Association, 101*(474), 460-471. doi:10.1198/016214505000000952

Markunas, C. A., Wilcox, A. J., Xu, Z. L., Joubert, B. R., Harlid, S., Panduri, V., . . . Taylor, J. A. (2016). Maternal Age at Delivery Is Associated with an Epigenetic Signature in Both Newborns and Adults. *Plos One, 11*(7). doi:ARTN e0156361 10.1371/journal.pone.0156361

Martin, E. M., & Fry, R. C. (2018). Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. *Annu Rev Public Health, 39*, 309-333. doi:10.1146/annurev-publhealth-040617-014629

Matsunaga, A., Hishima, T., Tanaka, N., Yamasaki, M., Yoshida, L., Mochizuki, M., . . . Hagiwara, S. (2014). DNA methylation profiling can classify HIV-associated lymphomas. *Aids, 28*(4), 503-510. doi:10.1097/Qad.0000000000000120

McInnes, T., Zou, D. H., Rao, D. S., Munro, F. M., Phillips, V. L., McCall, J. L., . . . Guilford, P. J. (2017). Genome-wide methylation analysis identifies a core set of hypermethylated genes in CIMP-H colorectal cancer. *Bmc Cancer, 17*. doi:ARTN 228 10.1186/s12885-017-3226-4

Meier, R., Graw, S., Usset, J., Raghavan, R., Dai, J., Chalise, P., . . . Koestler, D. (2016). An ensemble-based Cox proportional hazards regression framework for predicting survival in metastatic castration-resistant prostate cancer (mCRPC) patients. *F1000Res, 5*, 2677. doi:10.12688/f1000research.8226.1

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., . . . Irizarry, R. A. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods, 10*(10), 949-955. doi:10.1038/nmeth.2632

Moran, S., Arribas, C., & Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics, 8*(3), 389-399. doi:10.2217/epi.15.114

Nguyen, T., Li, G. E., Chen, H., Cranfield, C. G., McGrath, K. C., & Gorrie, C. A. (2018). Maternal E-Cigarette Exposure Results in Cognitive and Epigenetic Alterations in Offspring in a Mouse Model. *Chemical Research in Toxicology, 31*(7), 601-611. doi:10.1021/acs.chemrestox.8b00084

NIH, N. C. I. a. t. N. I. o. H. (2015). What Is Cancer? Retrieved from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

NIH, N. C. I. a. t. N. I. o. H. (2018). Cancer Statistics. Retrieved from https://www.cancer.gov/about-cancer/understanding/statistics

Omenn, G. S. (1997). Risk factors for lung cancer and for intervention effects in CARET, the Beta-Carotene and Retinol Efficacy Trial - Response. *Journal of the National Cancer Institute, 89*(22), 1723-1723. doi:DOI 10.1093/jnci/89.22.1723

Omenn, G. S., Goodman, G. E., Thornquist, M. D., Balmes, J., Cullen, M. R., Glass, A., . . . Hammar, S. (1996). Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. *New England Journal of Medicine, 334*(18), 1150-1155. doi:Doi 10.1056/Nejm199605023341802

Petrylak, D. P., Vogelzang, N. J., Budnik, N., Wiechno, P. J., Sternberg, C. N., Doner, K., . . . Fizazi, K. (2015). Docetaxel and prednisone with or without lenalidomide in chemotherapy-naive patients with metastatic castration-resistant prostate cancer (MAINSAIL): a randomised, double-blind, placebo-controlled phase 3 trial. *Lancet Oncology, 16*(4), 417-425. doi:10.1016/S1470-2045(15)70025-2

Pidsley, R., Wong, C. C. Y., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *Bmc Genomics, 14*. doi:Unsp 293 10.1186/1471-2164-14-293

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., . . . Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology, 17*. doi:ARTN 208 10.1186/s13059-016-1066-1

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics, 12*(8), 529-541. doi:10.1038/nrg3000

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y. F., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research, 43*(7). doi:ARTN e47 10.1093/nar/gkv007

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics, 6*(8), 597-610. doi:10.1038/nrg1655

Saadati, M., & Benner, A. (2014). Statistical challenges of high-dimensional methylation data. *Statistics in Medicine, 33*(30), 5347-5357. doi:10.1002/sim.6251

Saarela, O., Kulathinal, S., Arjas, E., & Laara, E. (2008). Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. *Statistics in Medicine, 27*(28), 5991-6008. doi:10.1002/sim.3416

Sakoda, L. C., Loomis, M. M., Doherty, J. A., Julianto, L., Barnett, M. J., Neuhouser, M. L., . . . Chen, C. (2012). Germ line variation in nucleotide excision repair genes and lung cancer risk in smokers. *Int J Mol Epidemiol Genet, 3*(1), 1-17.

Salim, A., Hultman, C., Sparen, P., & Reilly, M. (2009). Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics, 10*(1), 70-79. doi:10.1093/biostatistics/kxn016

Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika, 84*(2), 379-394. doi:DOI 10.1093/biomet/84.2.379

Samuelsen, S. O., Aring;nestad, H., & Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics, 34*(1), 103-119. doi:10.1111/j.1467-9469.2006.00552.x

Schallier, D., Decoster, L., Braeckman, J., Fontaine, C., & Degreve, J. (2012). Docetaxel in the treatment of metastatic castration-resistant prostate cancer (mCRPC): an observational study in a single institution. *Anticancer Res, 32*(2), 633-641.

Scher, H. I., Jia, X. Y., Chi, K., de Wit, R., Berry, W. R., Albers, P., . . . Heller, G. (2011). Randomized, Open-Label Phase III Trial of Docetaxel Plus High-Dose Calcitriol Versus Docetaxel Plus Prednisone for Patients With Castration-Resistant Prostate Cancer. *Journal of Clinical Oncology, 29*(16), 2191-2198. doi:10.1200/Jco.2010.32.8815

Schroder, M. S., Culhane, A. C., Quackenbush, J., & Haibe-Kains, B. (2011). survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics, 27*(22), 3206-3208. doi:10.1093/bioinformatics/btr511

Sestelo, M. (2017). A short course on Survival Analysis applied to the Financial Industry.

Siegmund, K. D. (2011). Statistical approaches for the analysis of DNA methylation microarray data. *Human Genetics, 129*(6), 585-595. doi:10.1007/s00439-011-0993-x

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics, 38*(2), 209-213. doi:10.1038/ng1706

Stoer, N. C., & Samuelsen, S. O. (2013). Inverse probability weighting in nested case-control studies with additional matching-a simulation study. *Statistics in Medicine, 32*(30), 5328-5339. doi:10.1002/sim.6019

Stoer, N. C., & Samuelsen, S. O. (2016). multipleNCC: Inverse Probability Weighting of Nested Case-Control Data. *R Journal, 8*(2), 5-18.

Syed, H., Jorgensen, A. L., & Morris, A. P. (2016). SurvivalGWAS_Power: a user friendly tool for power calculations in pharmacogenetic studies with "time to event" outcomes. *Bmc Bioinformatics, 17*. doi:ARTN 523 10.1186/s12859-016-1407-9

Tannock, I. F., Fizazi, K., Ivanov, S., Karlsson, C. T., Flechon, A., Skoneczna, I., . . . investigators, V. (2013). Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (VENICE): a phase 3, double-blind randomised trial. *Lancet Oncology, 14*(8), 760-768. doi:10.1016/S1470-2045(13)70184-0

Teschendorff, A. E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C., & Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Medicine, 4*. doi:ARTN 24 10.1186/gm323

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., & Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics, 29*(2), 189-196. doi:10.1093/bioinformatics/bts680

Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics, 19*(3), 129-147. doi:10.1038/nrg.2017.86

Therneau, T. M. (2015). A Package for Survival Analysis in S.

Thompson, J. A., Christensen, B. C., & Marsit, C. J. (2018). Methylation-to-Expression Feature Models of Breast Cancer Accurately Predict Overall Survival, Distant-Recurrence Free Survival, and Pathologic Complete Response in Multiple Cohorts. *Scientific Reports, 8*. doi:ARTN 5190 10.1038/s41598-018-23494-0

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological, 58*(1), 267-288.

Tsai, P. C., & Bell, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology, 44*(4), 1429-1441. doi:10.1093/ije/dyv041

Urdinguio, R. G., Torro, M. I., Bayon, G. F., Alvarez-Pitti, J., Fernandez, A. F., Redon, P., . . . Lurbe, E. (2016). Longitudinal study of DNA methylation during the first 5 years of life. *Journal of Translational Medicine, 14*. doi:ARTN 160 10.1186/s12967-016-0913-x

van Soest, R. J., Templeton, A. J., Vera-Badillo, F. E., Mercier, F., Sonpavde, G., Amir, E., . . . de Wit, R. (2015). Neutrophil-to-lymphocyte ratio as a prognostic biomarker for men with metastatic castration-resistant prostate cancer receiving first-line chemotherapy: data from two randomized phase III trials. *Annals of Oncology, 26*(4), 743-749. doi:10.1093/annonc/mdu569

Wang, S. (2011). Method to Detect Differentially Methylated Loci With Case-Control Designs Using Illumina Arrays. *Genetic Epidemiology, 35*(7), 686-694. doi:10.1002/gepi.20619

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society, 134*(630), 241-260. doi:10.1002/qj.210

WHO, W. H. O. (2007). Cancer control: Early detection - WHO guide for effective programmes. Retrieved from https://www.who.int/cancer/modules/Early%20Detection%20Module%203.pdf

WHO, W. H. O. (2018). Cancer. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/cancer

Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T., . . . Brown, R. (2013). Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer, 109*(6), 1394-1402. doi:10.1038/bjc.2013.496

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 73*, 3-36. doi:10.1111/j.1467-9868.2010.00749.x

Wu, H., Wang, C., & Wu, Z. J. (2015). PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics, 31*(2), 233-241. doi:10.1093/bioinformatics/btu640

Yang, X. J., Lay, F., Han, H., & Jones, P. A. (2010). Targeting DNA methylation for epigenetic therapy. *Trends in Pharmacological Sciences, 31*(11), 536-546. doi:10.1016/j.tips.2010.08.001

Zeilinger, S., Kuhnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., . . . Illig, T. (2013). Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *Plos One, 8*(5). doi:ARTN e63812 10.1371/journal.pone.0063812

Zhang, W. W., Spector, T. D., Deloukas, P., Bell, J. T., & Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology, 16*. doi:ARTN 14 10.1186/s13059-015-0581-9

Zhang, Y. H., Petropoulos, S., Liu, J. H., Cheishvili, D., Zhou, R., Dymov, S., . . . Szyf, M. (2018). The signature of liver cancer in immune cells DNA methylation. *Clinical Epigenetics, 10*. doi:ARTN 8 10.1186/s13148-017-0436-1

Zhang, Y. L. (2017). Recovery of weak signal in high dimensional linear regression by data perturbation. *Electronic Journal of Statistics, 11*(2), 3226-3250. doi:10.1214/17-Ejs1320

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418-1429. doi:10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). *Journal of the Royal Statistical Society Series B-Statistical Methodology, 67*, 768-768. doi:DOI 10.1111/j.1467-9868.2005.00527.x

Appendices

Appendix I - Integrated area under the curve (iAUC)

The integrated area under the curve (iAUC), defined by the DREAM website, is the averaged cumulative AUC over all time points (from 6 to 30 months).

Appendix II - Effect size boundary calculation

PDF of the beta distribution, $Beta(\alpha, \beta)$, with shape parameter $\alpha > 0$ and $\beta > 0$:

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

Where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Mean (μ) and variance (σ^2):

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{6.2.1}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{6.2.2}$$

Shape parameter as function of mean (μ) and variance (σ^2):

β as a function of α and μ from (6.2.1):

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\beta\mu = \alpha - \alpha\mu$$

$$\beta = \alpha\left(\frac{1}{\mu} - 1\right) \tag{6.2.3}$$

α as a function of μ and σ^2 (using (6.2.3) in (6.2.2)):

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\sigma^2 = \frac{\alpha^2\left(\frac{1}{\mu} - 1\right)}{\left(\alpha + \alpha\left(\frac{1}{\mu} - 1\right)\right)^2\left(\alpha + \alpha\left(\frac{1}{\mu} - 1\right) + 1\right)}$$

$$\sigma^2 = \frac{\left(\frac{1}{\mu} - 1\right)}{\left(1 + \left(\frac{1}{\mu} - 1\right)\right)^2 \left(\alpha + \alpha\left(\frac{1}{\mu} - 1\right) + 1\right)}$$

$$\alpha + \alpha\left(\frac{1}{\mu} - 1\right) + 1 = \frac{\left(\frac{1}{\mu} - 1\right)}{\sigma^2 \left(\frac{1}{\mu}\right)^2}$$

$$\alpha\left(\frac{1}{\mu}\right) = \frac{\left(\frac{1}{\mu} - 1\right)\mu^2}{\sigma^2} - 1$$

$$\alpha = \mu^2\left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) \tag{6.2.4}$$

β as a function of μ and σ^2 (using (6.2.4) in (6.2.3)):

$$\beta = \mu^2\left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right)\left(\frac{1}{\mu} - 1\right) \tag{6.2.5}$$

Relationship between μ and σ^2:

$$\alpha > 0$$

$$\mu^2\left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) > 0$$

$$\left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) > 0$$

$$\frac{1-\mu}{\sigma^2} > \frac{1}{\mu}$$

$$\mu - \mu^2 > \sigma^2 \tag{6.2.6}$$

Consider a modified mean $\mu^* = \mu + \Delta$, where the original mean μ was changed by Δ, while σ^2 remains unchanged. This modified mean μ^* must satisfy relationship (6.2.5), too:

$$\mu^* = \mu + \Delta$$

$$\mu^* - \mu^{*2} > \sigma^2$$

$$\mu + \Delta - (\mu + \Delta)^2 > \sigma^2$$

$$-\mu - \Delta + (\mu + \Delta)^2 < -\sigma^2$$

$$-\Delta + \mu^2 + 2\mu\Delta + \Delta^2 < \mu - \sigma^2$$

$$\Delta^2 + 2\Delta\left(\mu - \frac{1}{2}\right) < \mu - \mu^2 - \sigma^2$$

$$\Delta^2 + 2\Delta\left(\mu - \frac{1}{2}\right) + \left(\mu - \frac{1}{2}\right)^2 < \mu - \mu^2 - \sigma^2 + \left(\mu - \frac{1}{2}\right)^2$$

$$\left(\Delta + \left(\mu - \frac{1}{2}\right)\right)^2 < \mu - \mu^2 - \sigma^2 + \mu^2 - \mu + \frac{1}{4}$$

$$\Delta + \left(\mu - \frac{1}{2}\right) < \pm\sqrt{\frac{1}{4} - \sigma^2}$$

$$\Delta < \frac{1}{2} - \mu \pm \sqrt{\frac{1}{4} - \sigma^2} \qquad (6.2.7)$$

Hence, the imposed difference Δ is bounded by:

$$\frac{1}{2} - \mu - \sqrt{\frac{1}{4} - \sigma^2} < \Delta < \frac{1}{2} - \mu + \sqrt{\frac{1}{4} - \sigma^2} \qquad (6.2.8)$$

Appendix III - Additional figures

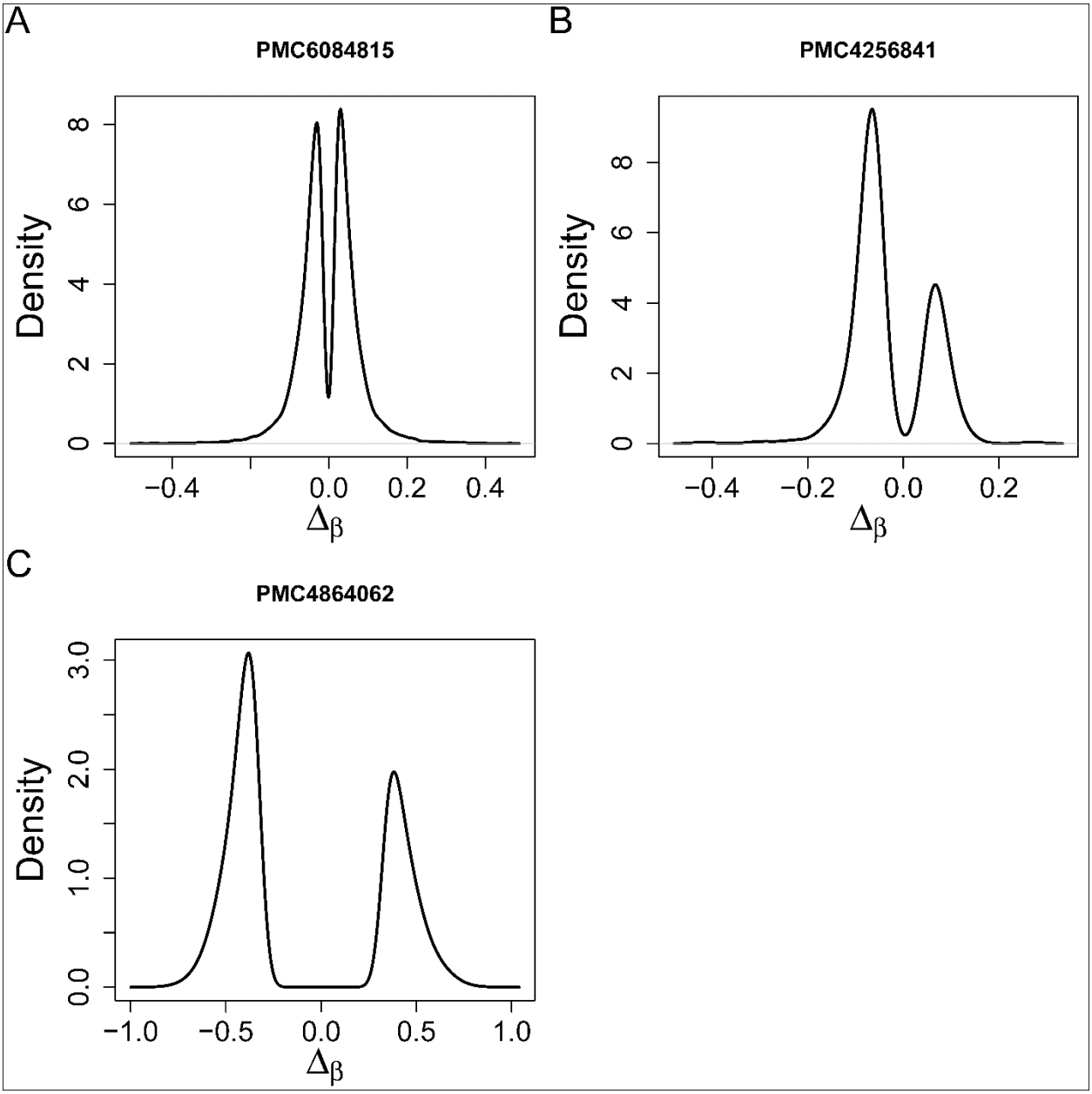


Figure S2.1: Empirical distributions of differences in DNAm for CpGs identified in different studies.

Density plots showing the distribution of differences in DNAm for CpGs identified in the following three studies: PMC6084815 (Hannon et al., 2018), PMC4256841 (Hall et al., 2014), and PMC4864062 (Moran, Arribas, & Esteller, 2016). The first study (PMC6084815) reported the differences in DNAm (Δ_β) of 20,001 differentially methylated CpGs (p-value < $1x10^{-8}$) comparing the additive genetic effects between monozygotic and dizygotic twins. The second study (PMC4256841) identified 470 differentially CpGs specific to the sex in human pancreatic islets. Difference in DNAm greater than 0.05 with a FDR less than 5% ($q < 0.05$) were provided. In the third study

(PMC4864062) two normal tissues were discriminated based on their methylation profile: colon mucosa (NC22A) and neurons (N229). Differences in DNAm greater than 0.33 were reported for 73,774 CpGs. It becomes apparent that simulating differences in DNAm from a truncated normal distribution, where values around zero (detection limit) are omitted, imitates observed values reasonably well. The truncation of the normal distribution is required due to the support of Δ_β ($-1 \leq \Delta_\beta \leq 1$). Even though the shown distributions exhibit some imbalance between positive and negative values, it is not necessary to preserve this potential imbalance as it will not affect the estimated power.

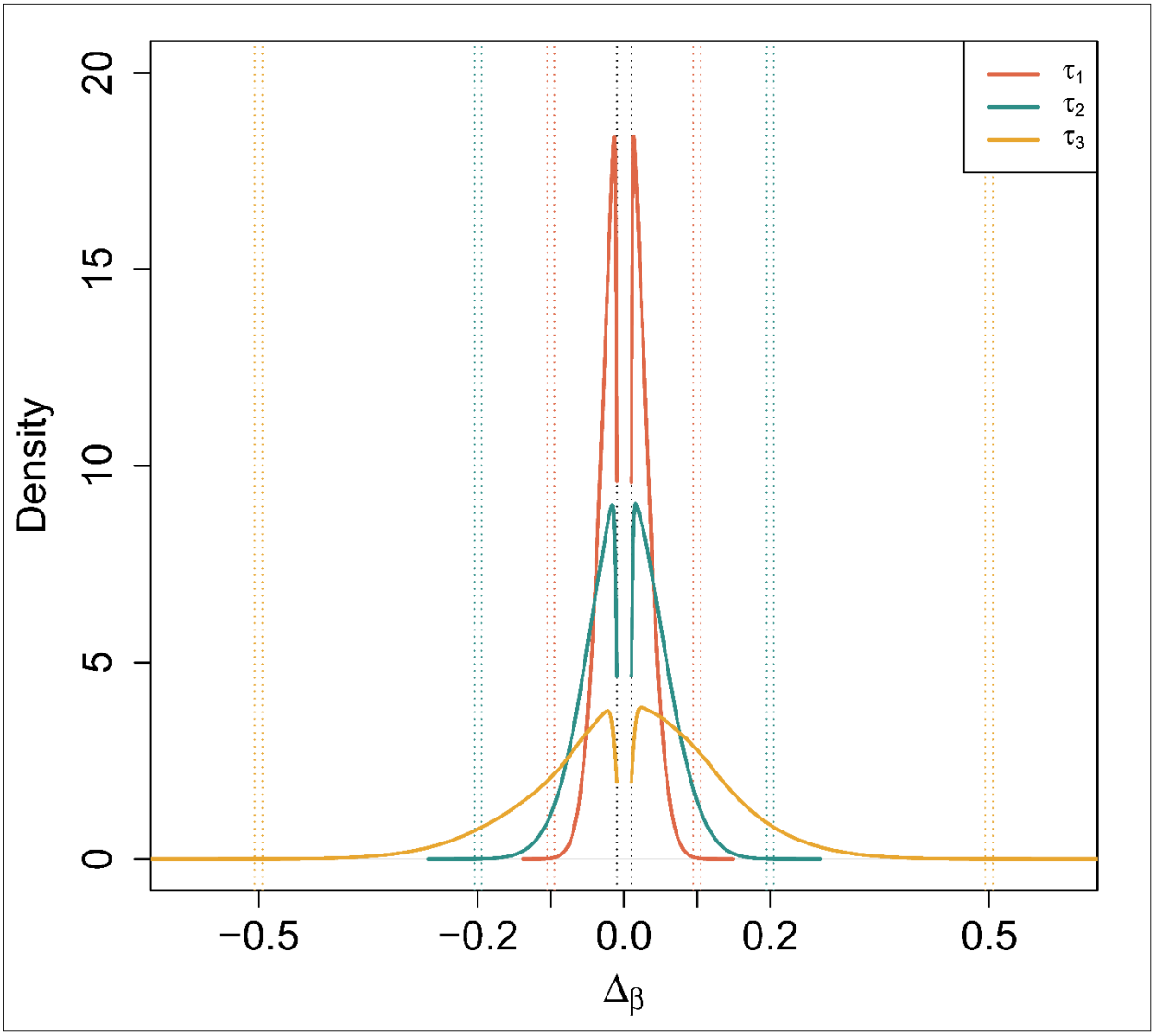


Figure S2.2: Distributions of simulated differences in DNAm (Δ_β) for different target effect sizes.

Δ_β is simulated using a truncated normal distribution $\Delta_{\beta,k} \sim N_k(0, \tau)$. To match the targeted maximal difference in DNAm, τ is stepwise adjusted until the 99.99th percentile of the absolute value of simulated $\Delta_{\beta,k}$'s falls within a range (vertical colored dotted lines) around the targeted maximal difference in DNAm. The range is equal the detection limit (vertical black dotted lines). Based on a default detection limit of 0.01, the 99.99th percentile needs to fall within a target effect size $\pm$ 0.005 for a τ to be accepted. The figure above shows densities of simulated $\Delta_{\beta,k}$'s for three effect sizes (0.1, 0.2, 0.5), and range for each effect size that the 99.99th percentile is required to fall in for τ to be accepted.

Appendix IV - Vignette

The pwrEWAS User's Guide

Stefan Graw, Devin C. Koestler

14 February 2019

Abstract

pwrEWAS is a user-friendly tool to estimate power in EWAS as a function of sample and effect size for two-group comparisons of DNAm (e.g., case vs control, exposed vs non-exposed, etc.). Detailed description of in-/outputs, instructions and an example, as well as interpretations of the example results are provided in the following vignette.

Package

pwrEWAS 0.99.1

Contents

Introduction

When designing an epigenome-wide association study (EWAS) to investigate the relationship between DNA methylation (DNAm) and some exposure(s) or phenotype(s), it is critically important to assess the sample size needed to detect a hypothesized difference with adequate statistical power. However, the complex and nuanced nature of DNAm data makes direct assessment of statistical power challenging. To circumvent these challenges and to address the outstanding need for a user-friendly interface for EWAS power evaluation, we have developed pwrEWAS. The current implementation of pwrEWAS accommodates power estimation for two-group comparisons of DNAm (e.g. case vs control, exposed vs non-exposed, etc.), where methylation assessment is carried out using the Illumina Human Methylation BeadChip technology. Power is calculated using a semi-parametric simulation-based approach in which DNAm data is randomly generated from beta-distributions using CpG-specific means and variances estimated from one of several different existing DNAm data sets, chosen to cover the most common tissue-types used in EWAS. In addition to specifying the tissue type to be used for DNAm profiling, users are required to specify the sample size, number of differentially methylated CpGs, effect size(s), target false discovery rate (FDR) and the number of simulated data sets, and have the option of selecting from several different statistical methods to perform differential methylation analyses. pwrEWAS reports the marginal power, marginal type I error rate, marginal FDR, and false discovery cost (FDC). The R-Shiny web inter-face allows for easy input of user-defined parameters and includes an advanced settings button that offers additional options pertaining to data generation and computation.

Installation

pwrEWAS can be installed from guthub with the following R code:

```
devtools::install_github("stefangraw/pwrEWAS")
library(pwrEWAS)
```

Usage

To execute the main pwrEWAS function the following to codes can be used. pwrEWAS allows the user to specify the effect size in one of two ways, by either providing a target maximal difference in methylation ("targetDelta"), or by providing the standard deviation of the simulated differnces ("deltaSD"). Only one of both arguments can be provided. If "targetDelta" is specified, pwrEWAS will automatically identify a standard deviation to simulate differences in methylation, such that the 99.99th percentile of the absolute value of simulated differences falls within a range around the targeted maximal difference in DNAm (see paper for additional details). If "deltaSD" is specified, pwrEWAS will simulate differences in methylation using the provided standard deviation (additional information provided in paper).

```
# providing the targeted maximal difference in DNAm
results_targetDelta = pwrEWAS(minTotSampleSize = 10,
                    maxTotSampleSize = 50,
                    SampleSizeSteps = 10,
                    NcntPer = 0.5,
                    targetDelta = c(0.2, 0.5),
                    J = 100,
                    targetDmCpGs = 10,
                    tissueType = "Adult (PBMC)",
                    detectionLimit = 0.01,
                    DMmethod = "limma",
                    FDRcritVal = 0.05,
                    core = 4,
                    sims = 50)

# providing the targeted maximal difference in DNAm
results_deltaSD = pwrEWAS(minTotSampleSize = 10,
                    maxTotSampleSize = 50,
                    SampleSizeSteps = 10,
                    NcntPer = 0.5,
                    deltaSD = c(0.02, 0.05),
                    J = 100,
                    targetDmCpGs = 10,
                    tissueType = "Adult (PBMC)",
                    detectionLimit = 0.01,
                    DMmethod = "limma",
                    FDRcritVal = 0.05,
                    core = 4,
                    sims = 50)
```

Input parameter

The following table provides a description of the input arguments:

Parameter	Description
minTotSampleSize	Lowest total sample sizes to be considered
maxTotSampleSize	Highest total sample sizes to be considered
SampleSizeSteps	Steps with which total sample size increases from minTotSampleSize to maxTotSampleSize
NcntPer	Rate by which the total sample size is split into groups (0.5 corresponds to a balanced study; rate for group 2 is equal to 1 rate of group 1)
targetDelta	Standard deviations of the simulated differences is automatically determined such that the 99%til of the simulated differences are within a range around the provided values
deltaSD	Differences in methylation will be simulated using provided standard deviation
J	Number of CpG site that will simulated and tested (increasing Number of CpGs tested will require increasing RAM (memory))
targetDmCpGs	Target number of CpGs simulated with meaningful differences (differences greater than detection limit)
tissueType	Heterogeneity of different tissue types can have effects on the results. Please select your tissue type of interest or one you believe is the closest
detectionLimit	Limit to detect changes in methylation. Simulated differences below the detection limit will not be consider as meaningful differentially methylated CpGs
DMmethod	Method used to perform differential methylation analysis
FDRcritVal	Critical value to control the False Discovery Rate (FDR) using the Benjamini and Hochberg method
core	Number of cores used to run multiple threads. Ideally, the number of different total samples sizes multiplied by the number of effect sizes should be a multiple (m) of the number of cores (#sampleSizes * #effectSizes = m * #threads). An increasing number of threads will require an increasing amount of RAM (memory)
sims	Number of repeated simulation/simulated data sets under the same conditions for consistent results

Output parameter

Running pwrEWAS will result in an object with the following four attributes: meanPower, powerArray, deltaArray, and metric. The first attribute "meanPower"is a 2D matrix with empirically estimated marginal mean power for sample sizes and target Δ_βs (averaged over simulated data sets). The second attribute"powerArray"provides the full set of empirically estimated marginal power for sample sizes and target Δ_βs for each simulated data sets in a 3D matrix. The third attribute"deltaArray"contains a 3D matrix with simulated Δ_βs for sample sizes, target Δ_β, and simulated data sets. The last attribute"metric" contains 2D matrices with the marginal type I error rate (marTypeI), power in the classical sense (classicalPower),

actual FDR (FDR), False Discovery Cost (FDC), and probabilities of identifying at least one true positive in table format, where sample sizes are shown as rows and effect sizes are columns. Examples results can be found in the example section.

Runtime

In general, the computational complexity of pwrEWAS depends on four major components: (1) assumed number and magnitude of sample size(s), (2) number of target Δ_β's (effect sizes), (3) number of CpGs tested, and (4) number of simulated data sets. To enhance the computational efficiency, pwrEWAS allows users to process simulations in parallel. While (1) and (2) are usually dictated by the study to be conducted, (3) and (4) can be modified to either increase the precision of power estimates (increased run time) or reduce the computational burden (decreased precision of estimates). The following table provides the run time of pwrEWAS for different combinations of sample sizes and effect sizes. In all scenarios presented the number of tested CpGs was assumed to be 100,000, number of simulated data sets was 50, and the method to perform the differential methylation analysis as limma. A total of 6 clusters/threads were used.

Total sample size Effect size Δ_β	0.1	0.1, 0.2	0.1, 0.3, 0.5
10	2min 21sec	3min 11sec	3min 50sec
100	6min 22sec	7min 39sec	8min 33sec
500	24min 43sec	27min 36sec	29min 22sec
10-100 (increments of 10)	9min 40sec	16min 34sec	23min 44sec
300-500 (increments of 100)	27min 58sec	30min 01sec	52min 00sec

Example

Running pwrEWAS

Running pwrEWAS by providing target maximal difference in methylation:

```
library(pwrEWAS)
set.seed(1234)
results_targetDelta = pwrEWAS(minTotSampleSize = 20,
                    maxTotSampleSize = 260,
                    SampleSizeSteps = 40,
                    NcntPer = 0.5,
                    targetDelta = c(0.02, 0.10, 0.15, 0.20),
                    J = 100000,
                    targetDmCpGs = 2500,
                    tissueType = "Blood adult",
                    detectionLimit = 0.01,
                    DMmethod = "limma",
                    FDRcritVal = 0.05,
                    core = 4,
                    sims = 50)
```

Running pwrEWAS by providing standard deviation of difference in methylation:

```
library(pwrEWAS)
set.seed(1234)
results_deltaSD = pwrEWAS(minTotSampleSize = 20,
                    maxTotSampleSize = 260,
                    SampleSizeSteps = 40,
                    NcntPer = 0.5,
                    deltaSD = c(0.00390625, 0.02734375, 0.0390625, 0.052734375),
                    J = 100000,
                    targetDmCpGs = 2500,
                    tissueType = "Blood adult",
                    detectionLimit = 0.01,
                    DMmethod = "limma",
                    FDRcritVal = 0.05,
                    core = 4,
                    sims = 50)
```

If pwrEWAS is executed with providing target maximal difference, first τ will be determined. The beginning and finish of this process will be printed with time stamps (see below for an example). If the standard deviation of difference is provided, this step will be skipped. \ Next, pwrEWAS will run the simulations to empirically estimate power. pwrEWAS will indicate when the simulations are started. To monitor the process pwrEWAS will display a process bar. pwrEWAS will print a statement including a time stamps once finished (see below for an example).

```
## [2019-02-12 18:40:23] Finding tau...done [2019-02-12 18:42:53]
## [1] "The following taus were chosen: 0.00390625, 0.02734375, 0.0390625, 0.052734375"
## [2019-02-12 18:42:53] Running simulation
## |============================================================| 100%
## [2019-02-12 18:42:53] Running simulation ... done [2019-02-12 19:27:03]
```

Outputs

Running pwrEWAS will result in an object, that stores the following four attributes:

```
attributes(results_targetDelta)
## $names
## [1] "meanPower" "powerArray" "deltaArray" "metric"
```

meanPower

The primary results will be provided in the attribute ''meanPower''. It is essentially a summary of the attribute''powerArray''. meanPower will be provide a 7x4 table with the average power by total sample size as rows (here 20-260 patients with increments of 40) and by target Δ_β, if''targetDelta''was provided, or''$SD(\Delta_\beta)$'', if deltaSD was provided, as columns (here targetDelta was provided as: 0.02, 0.1, 0.15, 0.2):

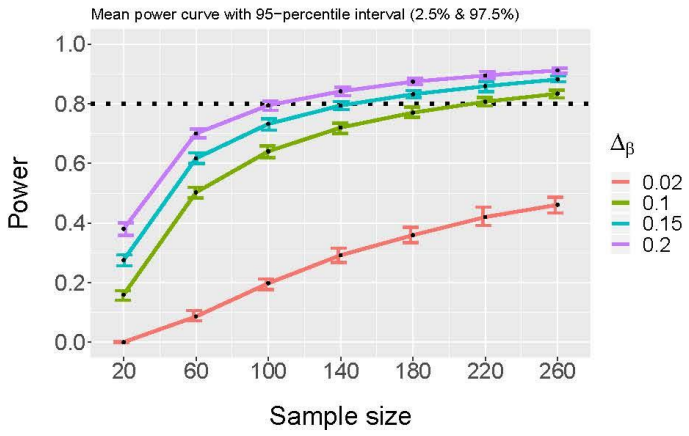
```
dim(results_targetDelta$meanPower)
## [1] 7 4
```

```
print(results_targetDelta$meanPower)
##             0.02       0.1      0.15       0.2
## 20  0.0005415101 0.1596165 0.2758319 0.3801848
## 60  0.0863276853 0.5026172 0.6166725 0.7001472
## 100 0.1978966524 0.6402466 0.7322670 0.7947203
## 140 0.2919669218 0.7201027 0.7940429 0.8414375
## 180 0.3592038789 0.7700964 0.8317636 0.8739818
## 220 0.4201022535 0.8068096 0.8588536 0.8945975
## 260 0.4609956067 0.8338529 0.8816222 0.9117306
```

powerArray

The attribute "powerArray"should primarily be used to create a power plot but can also be used to investigate the power results for the individual simulations. pwrEWAS includes a function"pwrEWAS_powerPlot"that will create a power plot, where power (y-axis) is shown as a function of sample sizes (x-axis) for different effect sizes (color coded). For each sample size, the mean power as well as the 95%tile interval (2.5% and 97.5%) is shown."sd"should be set to"FALSE"if"targetDelta"was specified in pwrEWAS, and"TRUE"if"deltaSD" was specified in pwrEWAS.

```
dim(results_targetDelta$powerArray) # simulations x sample sizes x effect sizes
## [1] 50  7  4
pwrEWAS_powerPlot(results_targetDelta$powerArray, sd = FALSE)
```



deltaArray

The third attribute "deltaArray"contains the simulated differences in mean DNAm. Each Δ_β is drawn from a truncated normal, where either the standard devation is provided ("deltaSD") or automatically determined based on the user-specified target Δ_β ("targetDelta") and the expected number of differentially methylated CpGs ("targetDmCpGs"). To automatically determined the standard devation, it is adjusted stepwise until the 99.99th percentile of the

absolute value of simulated Δ_βs falls within a range around the targeted maximal difference in DNAm (see paper for additional details). The maximal value of Δ_β can exceed the user-specified target Δ_β, but about 99.99% of simulated differences will be below user-specified target Δ_β (as seen below):

```
# maximum value of simulated differences by target value
lapply(results_targetDelta$deltaArray, max)
## $`0.02`
## [1] 0.02095302
##
## $`0.1`
## [1] 0.1265494
##
## $`0.15`
## [1] 0.2045638
##
## $`0.2`
## [1] 0.2458416

# percentage of simulated differences to be within the target range
mean(results _ targetDelta$deltaArray[[1]] < 0.02)
## [1] 0.9999999
mean(results _ targetDelta$deltaArray[[2]] < 0.10)
## [1] 0.9998882
mean(results _ targetDelta$deltaArray[[3]] < 0.15)
## [1] 0.9999386
mean(results _ targetDelta$deltaArray[[4]] < 0.20)
## [1] 0.9999539
```

To get a better understanding of how the differences in mean DNAm are distributed, pwrEWAS provides a density plot, where the distribution of simulated differences in mean DNAm is plotted by target differences in DNAm (Δ_β). The color theme matches the colors of the power plot. Simulated differences within the detection limit around zero are removed, as they are here not defined as meaningful differences. "sd"should be set to"FALSE"if"targetDelta"was specified in pwrEWAS, and"TRUE"if"deltaSD" was specified in pwrEWAS.

```
pwrEWAS deltaDensity(results targetDelta$deltaArray, detectionLimit = 0.01, sd = FALSE)
```

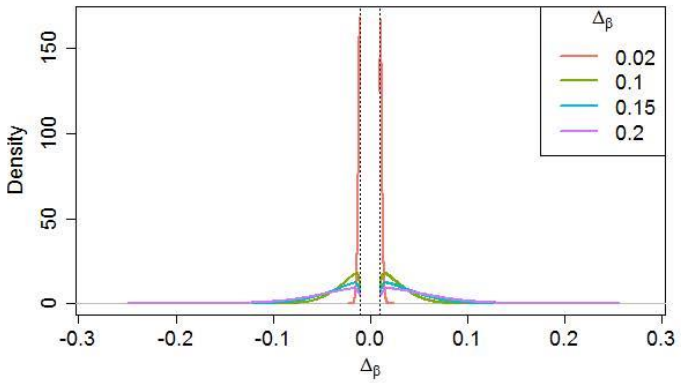


Figure 1:

In the figure above, the densities are very compress, because the first effect size is clearly different from the other three. The following code will provide the figure after removing the first effect size:

```
temp = results targetDelta$deltaArray
temp[[1]] = NULL
pwrEWAS deltaDensity(temp, detectionLimit = 0.01, sd = FALSE)
```

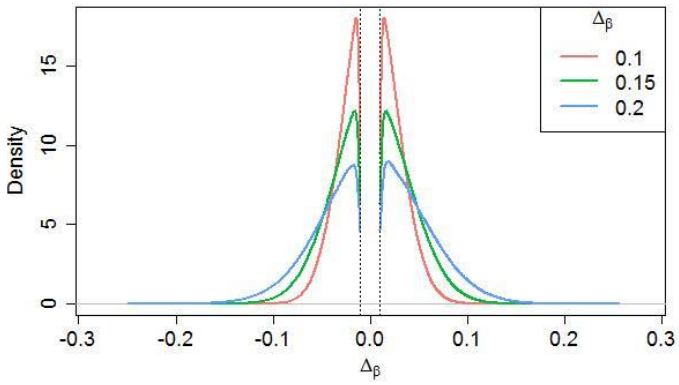


Figure 2:

metric

The fourth attribute "metric"contains tables on marginal type I error rate ("marTypeI"), power in the classical sense (classicalPower), actual FDR (FDR), False Discovery Cost (FDC, see paper for additional details), and probabilities of identifying at least one true positive, for each sample size and effect size combination:

```
results targetDelta$metric
## $marTypeI
##             0.02          0.1         0.15          0.2
## 20  0.0000000000 0.0001927742 0.0003533407 0.0004575820
## 60  0.0003435644 0.0006813394 0.0008155174 0.0009199059
## 100 0.0011254329 0.0008494543 0.0009544978 0.0010784126
## 140 0.0023155015 0.0009987010 0.0011007120 0.0011301504
## 180 0.0031646165 0.0010936404 0.0011537869 0.0011709668
## 220 0.0043017549 0.0011711588 0.0011688609 0.0011961111
## 260 0.0050251766 0.0011825572 0.0012256528 0.0012703139
##
## $classicalPower
##          0.02       0.1      0.15       0.2
## 20  0.0000230 0.1140913 0.2188748 0.3211014
## 60  0.0072840 0.3665948 0.4969722 0.5948422
## 100 0.0243978 0.4749589 0.5952528 0.6816387
## 140 0.0447472 0.5386504 0.6500822 0.7252384
## 180 0.0650878 0.5859314 0.6852212 0.7575621
## 220 0.0848528 0.6181004 0.7131985 0.7789187
## 260 0.1031464 0.6445875 0.7373262 0.7980121
##
## $FDR
```

```
##            0.02        0.1       0.15        0.2
## 20  0.000000000 0.04402833 0.04704810 0.04431709
## 60  0.004517889 0.04837274 0.04781140 0.04793149
## 100 0.004418029 0.04660417 0.04675174 0.04899762
## 140 0.004953809 0.04824695 0.04925629 0.04831297
## 180 0.004662894 0.04856488 0.04900589 0.04792516
## 220 0.004844068 0.04925471 0.04774217 0.04763118
## 260 0.004668246 0.04777146 0.04839480 0.04928052
##
## $FDC
##            0.02        0.1       0.15        0.2
## 20  0.00000000 0.04638179 0.04959881 0.04652390
## 60  0.03043364 0.05207772 0.05098151 0.05083022
## 100 0.04361648 0.05082257 0.05032126 0.05243835
## 140 0.06076271 0.05345012 0.05358616 0.05197136
## 180 0.06718904 0.05446132 0.05373615 0.05186793
## 220 0.07820442 0.05588333 0.05261712 0.05173305
## 260 0.08348369 0.05445357 0.05369060 0.05387600
##
## $probTP
##     0.02 0.1 0.15 0.2
## 20   0.4   1    1   1
## 60   1.0   1    1   1
## 100  1.0   1    1   1
## 140  1.0   1    1   1
## 180  1.0   1    1   1
## 220  1.0   1    1   1
## 260  1.0   1    1   1
```

Interpretation

To detect differences up to 10%, 15% and 20% in CpG-specific methylation across 2,500 CpGs with at least 80% power, we would need about 220, 180 and 140 total subjects, respectively. As expected, 80% power was not achieved for a difference in DNAm up to 2% for the selected total sample size range. However, it can be observed that the probability of detecting at least one CpG out of the 2500 differentially methylated CpGs is about 40% for 20 total patients and virtually 100% for 60 and more total patients.

SessionInfo

```
toLatex(sessionInfo())
```

- R version 3.5.2 (2018-12-20), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`

- Running under: `Windows 7 x64 (build 7601) Service Pack 1`
- Matrix products: default
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: BiocStyle 2.10.0, foreach 1.4.4, pwrEWAS 0.99.1, shinyBS 0.61
- Loaded via a namespace (and not attached): annotate 1.60.0, AnnotationDbi 1.44.0, assertthat 0.2.0, backports 1.1.3, Biobase 2.42.0, BiocGenerics 0.28.0, BiocManager 1.30.4, bit 1.1-14, bit64 0.9-7, bitops 1.0-6, blob 1.1.1, bookdown 0.9, callr 3.1.1, cli 1.0.1, codetools 0.2-16, colorspace 1.4-0, compiler 3.5.2, crayon 1.3.4, curl 3.3, DBI 1.0.0, desc 1.2.0, devtools 2.0.1, digest 0.6.18, doParallel 1.0.14, evaluate 0.12, fs 1.2.6, genefilter 1.64.0, ggplot2 3.1.0, glue 1.3.0, grid 3.5.2, gtable 0.2.0, htmltools 0.3.6, httpuv 1.4.5.1, IRanges 2.16.0, iterators 1.0.10, knitr 1.21, later 0.8.0, lattice 0.20-38, lazyeval 0.2.1, limma 3.38.3, magrittr 1.5, Matrix 1.2-15, memoise 1.1.0, mime 0.6, munsell 0.5.0, parallel 3.5.2, pillar 1.3.1, pkgbuild 1.0.2, pkgconfig 2.0.2, pkgload 1.0.2, plyr 1.8.4, prettyunits 1.0.2, processx 3.2.1, promises 1.0.1, ps 1.3.0, R6 2.3.0, Rcpp 1.0.0, RCurl 1.95-4.11, remotes 2.0.2, rlang 0.3.1, rmarkdown 1.11, rprojroot 1.3-2, RSQLite 2.1.1, rstudioapi 0.9.0, S4Vectors 0.20.1, scales 1.0.0, sessioninfo 1.1.1, shiny 1.2.0, splines 3.5.2, stats4 3.5.2, stringi 1.2.4, stringr 1.4.0, survival 2.43-3, tibble 2.0.1, tools 3.5.2, usethis 1.4.0, withr 2.1.2, xfun 0.4, XML 3.98-1.17, xtable 1.8-3, yaml 2.2.0

References

Appendix V - List of potential weighting functions

Function based on density of Normal Density:

$$f\left(\Delta_k^{(m)}\right) = \frac{1}{2}\left(sign\left(\Delta_k^{(m)}\right) * \left(1 - \frac{f_N\left(\Delta_k^{(m)}\right)}{f_N(0)}\right) + 1\right)$$

where $f_N(\cdot)$ is the density of the Normal distribution with mean $\mu = 0$ and a standard deviation that controls the steepness.

Sigmoid function:

$$f\left(\Delta_k^{(m)}\right) = \frac{1}{1 + \exp(-\tau\Delta_k^{(m)})}$$

where τ controls the steepness.

Cubic function:

$$f\left(\Delta_k^{(m)}\right) = \frac{{\Delta_k^{(m)}}^3 + 1}{2}$$

Linear function:

$$\left(\Delta_k^{(m)}\right) = \frac{1}{2}\Delta_k^{(m)} + 0.5$$

The curves of these function can are visualized in Figure S3.1:

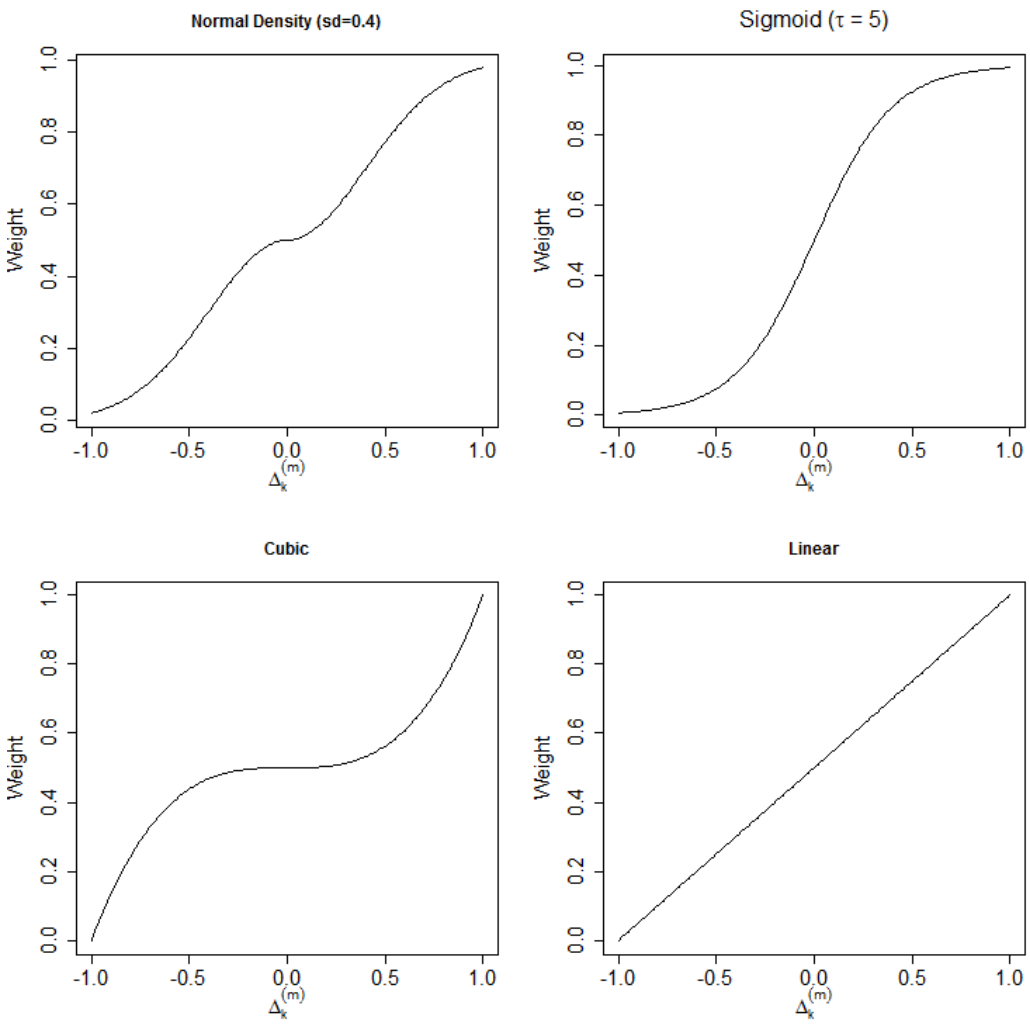


Figure S3.1: Weight function curves

Appendix VI - Ratio of gamma distributed signals follows beta distribution

Assume the methylated signal *M* and unmethylated signal *U* follow a gamma distribution with equal scale parameter:

$$M \sim Gamma(a, \theta)$$

$$U \sim Gamma(b, \theta)$$

Their joint distribution is given by:

$$f_{m,u}(m,u) = \frac{1}{\Gamma(a)\theta^a} m^{a-1} e^{-\frac{m}{\theta}} \frac{1}{\Gamma(b)\theta^b} u^{b-1} e^{-\frac{u}{\theta}}$$

$$= \frac{1}{\Gamma(a)\Gamma(b)\theta^{a+b}} m^{a-1} u^{b-1} e^{-\frac{1}{\theta}(m+u)}$$

Let:

$$x = u + m$$

$$y = \frac{m}{m+u}$$

Rearranging for *m* and *u*:

$$m = xy$$

$$u = x - xy$$

The Jacobian can be calculated as:

$$J = \begin{vmatrix} y & x \\ 1-y & -x \end{vmatrix} = -xy - (x - xy) = -x$$

The transformation is then given as:

$$f_{x,y}(x,y) = f_{m,u}(m,u)|J|$$

$$= f_{m,u}(m = xy, u = x - xy)|-x|$$

$$= \frac{1}{\Gamma(a)\Gamma(b)\theta^{a+b}}(xy)^{a-1}(x - xy)^{b-1}e^{-\frac{1}{\theta}(xy+x-xy)}x$$

$$= \frac{1}{\Gamma(a)\Gamma(b)\theta^{a+b}}x^{a-1}y^{a-1}(1-y)^{b-1}x^{b-1}e^{-\frac{x}{\theta}}x$$

$$= \frac{1}{\Gamma(a)\Gamma(b)\theta^{a+b}}x^{a-1+b-1+1}y^{a-1}(1-y)^{b-1}e^{-\frac{x}{\theta}}\frac{\Gamma(a+b)}{\Gamma(a+b)}$$

$$= \frac{1}{\Gamma(a+b)\theta^{a+b}}x^{a+b-1}e^{-\frac{x}{\theta}}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1}$$

$$= f(x)f(y)$$

By the factorization theorem, *x* and *y* are independent.

$$f(x) = \frac{1}{\Gamma(a+b)\theta^{a+b}}x^{a+b-1}e^{-\frac{x}{\theta}} \sim Gamma(\alpha+\beta,\theta)$$

$$f\left(y = \frac{m}{m+u}\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(1-y)^{b-1}y^{a-1} \sim Beta(a,b)$$

When *M* and *U* are gamma distributed random variables with equal scale parameter, then their ratio $\left(y = \frac{\mathrm{m}}{\mathrm{m+u}}\right)$ follows a beta distribution.

Appendix VII - Generation of correlated beta values

Let's first show that the marginal distribution of a Dirichlet distribution is beta distributed:

Given:

$$\vec{X} \sim Dirichlet(\vec{\alpha})$$

$$\vec{X} = [X_1, X_2, \dots, X_P]$$

$$\sum_{i=1}^{P} x_i = 1$$

$$\vec{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_P]$$

$$\alpha_i > 0$$

$$\sum_{i=1}^{P} \alpha_i = A$$

$$f(\vec{x}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{P} x_i^{\alpha_i - 1}$$

$$B(\vec{\alpha}) = \frac{\prod_{i=1}^{P} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{P} \alpha_i)}$$

We can write the joint density of $x_1, x_2, \dots, x_P$ as:

$$f(\vec{x}) = f\left(x_1, x_2, \dots, x_{P-1}, \left(1 - \sum_{i=1}^{P-1} x_i\right)\right) = f(x_1, x_2, \dots, x_{P-1})$$

Using: $f(x, y) = f(x|y)f(y)$

$$= f(x_1, x_2, \dots, x_{P-2}) f(x_{P-1} | x_1, x_2, \dots, x_{P-2})$$

$$= f(x_1, x_2, \dots, x_{P-3}) f(x_{P-2} | x_1, x_2, \dots, x_{P-3}) f(x_{P-1} | x_1, x_2, \dots, x_{P-2})$$

...

$$= f_1(x_1) f_2(x_2|x_1) f_3(x_3|x_1, x_2) \dots f_{P-1}(x_{P-1}|x_1, \dots, x_{P-2})$$

Expressions for each term individually:

$$f_1(x_1) = \frac{\Gamma(\alpha_1 + A - \alpha_1)}{\Gamma(\alpha_1)\Gamma(A - \alpha_1)} x_1^{\alpha_1 - 1}(1 - x_1)^{A - \alpha_1 - 1}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A - \alpha_1)} x_1^{\alpha_1 - 1}(1 - x_1)^{A - \alpha_1 - 1}$$

$$f_2(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)}$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + A - \alpha_1 - \alpha_2) x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} \cdot$$

$$\frac{\Gamma(\alpha_1)\Gamma(A - \alpha_1)}{\Gamma(\alpha_1 + A - \alpha_1) x_1^{\alpha_1 - 1}(1 - x_1)^{A - \alpha_1 - 1}}$$

$$= \frac{\Gamma(A) x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} \frac{\Gamma(\alpha_1)\Gamma(A - \alpha_1)}{\Gamma(A) x_1^{\alpha_1 - 1}(1 - x_1)^{A - \alpha_1 - 1}}$$

$$= \frac{\Gamma(A - \alpha_1)}{\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} \frac{x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}}{(1 - x_1)^{A - \alpha_1 - 1}}$$

Analogously:

$$f_3(x_3|x_1x_2) = \frac{\Gamma(A-\alpha_1-\alpha_2)}{\Gamma(\alpha_3)\Gamma(A-\alpha_1-\alpha_2-\alpha_3)} \frac{x_3^{\alpha_3-1}(1-x_1-x_2-x_3)^{A-\alpha_1-\alpha_2-\alpha_3-1}}{(1-x_1-x_2)^{A-\alpha_1-\alpha_2-1}}$$

$$f_{P-1}(x_{P-1}|x_1x_2 \dots x_{P-1}) = \frac{\Gamma(A-\alpha_1-\cdots-\alpha_{P-2})}{\Gamma(\alpha_{P-1})\Gamma(A-\alpha_1-\cdots-\alpha_{P-1})} \cdot$$

$$\frac{x_{P-1}^{\alpha_{P-1}-1}(1-x_1-\cdots-x_{P-1})^{A-\alpha_1-\cdots-\alpha_{P-1}-1}}{(1-x_1-\cdots-x_{P-2})^{A-\alpha_1-\cdots-\alpha_{P-2}-1}}$$

$$= \frac{\Gamma(A-\alpha_1-\cdots-\alpha_{P-2})}{\Gamma(\alpha_{P-1})\Gamma(\alpha_P)} \frac{x_{P-1}^{\alpha_{P-1}-1}x_P^{\alpha_P-1}}{(1-x_1-\cdots-x_{P-2})^{\alpha_{P-1}+\alpha_P-1}}$$

Plugging individual terms back in:

$$f(x) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1} \cdot$$

$$= \frac{\Gamma(A-\alpha_1)}{\Gamma(\alpha_2)\Gamma(A-\alpha_1-\alpha_2)} \frac{x_2^{\alpha_2-1}(1-x_1-x_2)^{A-\alpha_1-\alpha_2-1}}{(1-x_1)^{A-\alpha_1-1}} \cdot$$

$$\frac{\Gamma(A-\alpha_1-\alpha_2)}{\Gamma(\alpha_3)\Gamma(A-\alpha_1-\alpha_2-\alpha_3)} \frac{x_3^{\alpha_3-1}(1-x_1-x_2-x_3)^{A-\alpha_1-\alpha_2-\alpha_3-1}}{(1-x_1-x_2)^{A-\alpha_1-\alpha_2-1}} \cdot$$

…

$$\frac{\Gamma(A-\alpha_1-\cdots-\alpha_{P-2})}{\Gamma(\alpha_{P-1})\Gamma(\alpha_P)} \frac{x_{P-1}^{\alpha_{P-1}-1}x_P^{\alpha_P-1}}{(1-x_1-\cdots-x_{P-2})^{\alpha_{P-1}+\alpha_P-1}}$$

Second part of the denominator cancels with the nominator of the successor:

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)} x_1^{\alpha_1 - 1}(1 - x_1)^{A - \alpha_1 - 1} \cdot$$

$$\frac{1}{\Gamma(\alpha_2)} \frac{x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}}{(1 - x_1)^{A - \alpha_1 - 1}} \cdot$$

$$\frac{1}{\Gamma(\alpha_3)} \frac{x_3^{\alpha_3 - 1}(1 - x_1 - x_2 - x_3)^{A - \alpha_1 - \alpha_2 - \alpha_3 - 1}}{(1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}} \cdot$$

...

$$\frac{1}{\Gamma(\alpha_{P-1})\Gamma(\alpha_P)} \frac{x_{P-1}^{\alpha_{P-1} - 1} x_P^{\alpha_P - 1}}{(1 - x_1 - \cdots - x_{P-2})^{\alpha_{P-1} + \alpha_P - 1}}$$

Last term of each pdf cancels with the second denominator of the successor:

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)} x_1^{\alpha_1 - 1} \cdot$$

$$\frac{1}{\Gamma(\alpha_2)} x_2^{\alpha_2 - 1} \cdot$$

$$\frac{1}{\Gamma(\alpha_3)} x_3^{\alpha_3 - 1} \cdot$$

...

$$\frac{1}{\Gamma(\alpha_{P-1})\Gamma(\alpha_P)} x_{P-1}^{\alpha_{P-1} - 1} x_P^{\alpha_P - 1}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)} x_1^{\alpha_1 - 1} \frac{1}{\Gamma(\alpha_2)} x_2^{\alpha_2 - 1} \frac{1}{\Gamma(\alpha_3)} x_3^{\alpha_3 - 1} \ \ldots \ \frac{1}{\Gamma(\alpha_{P-1})} x_{P-1}^{\alpha_{P-1} - 1} \frac{1}{\Gamma(\alpha_P)} x_P^{\alpha_P - 1}$$

$$= \frac{\Gamma(\sum_{i=1}^{P} \alpha_1)}{\prod_{i=1}^{P} \Gamma(\alpha_i)} \prod_{i=1}^{P} x_i^{\alpha_i - 1}$$

If $X_i \sim beta(\alpha_1, A - \alpha_i)$ the joint distribution $\vec{X} \sim dirichlet(\vec{\alpha})$.

Alternatively, the Dirichlet aggregation property can be used:

If

$$\vec{X} = (X_1, \dots, X_P) \sim Dir(\alpha_1, \dots, \alpha_P)$$

then, if the random variables with subscripts i and j are dropped from the vector and replaced by their sum

$$\overrightarrow{X'} = \left(X_1, \dots, X_i + X_j, \dots, X_P\right) \sim Dir\left(\alpha_1, \dots, \alpha_i + \alpha_j, \dots, \alpha_P\right)$$

This aggregation property may be used to derive the marginal distribution of X_i. Let's pick X_1 and X_2 as an example:

$$f(x_1, x_2|\alpha_1, \alpha_2, A) = \frac{\Gamma(\alpha_1 + \alpha_2 + A - \alpha_1 - \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1}$$

Integrating out x_2

$$f(x_1|\alpha_1, \alpha_2, A) = \int_0^{1 - x_1} f(x_1, x_2|\alpha_1, \alpha_2, A) dx_2$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} \int_0^{1 - x_1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{A - \alpha_1 - \alpha_2 - 1} dx_2$$

Change of variables:

$$u = \frac{x_2}{1 - x_1} = g(x_2)$$

$$x_2 = (1 - x_1)u$$

$$dx_2 = (1 - x_1)du$$

$$f(x_1|\alpha_1, \alpha_2, A) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} \cdot$$

$$\int_{\frac{0}{1-x_1}}^{\frac{1-x_1}{1-x_1}} \big((1 - x_1)u\big)^{\alpha_2 - 1} (1 - x_1 - (1 - x_1)u)^{A - \alpha_1 - \alpha_2 - 1} (1 - x_1) du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} \cdot$$

$$\int_0^1 (1 - x_1)^{\alpha_2 - 1} u^{\alpha_2 - 1} \big((1 - x_1)(1 - u)\big)^{A - \alpha_1 - \alpha_2 - 1} (1 - x_1) du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} (1 - x_1)^{\alpha_2 - 1 + A - \alpha_1 - \alpha_2 - 1 + 1} \cdot$$

$$\int_0^1 u^{\alpha_2 - 1} (1 - u)^{A - \alpha_1 - \alpha_2 - 1} du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A - \alpha_1 - \alpha_2)} x_1^{\alpha_1 - 1} (1 - x_1)^{A - \alpha_1 - 1} \cdot$$

$$\int_0^1 u^{\alpha_2 - 1} (1 - u)^{A - \alpha_1 - \alpha_2 - 1} du$$

By the property that a beta pdf integrates to 1:

$$\int_0^1 u^{a-1}(1-u)^{b-1}du = \frac{\Gamma(b)\Gamma(b)}{\Gamma(a+b)}$$

$$f(x_1|\alpha_1,\alpha_2,A) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(A-\alpha_1-\alpha_2)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1}$$

$$\frac{\Gamma(\alpha_2)\Gamma(A-\alpha_1-\alpha_2)}{\Gamma(\alpha_2+A-\alpha_1-\alpha_2)}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1}$$

$$= f(x_1|\alpha_1,A)$$

$$X_1 \sim beta(\alpha_1, A-\alpha_1)$$

If $\vec{X} \sim dirichlet(\vec{\alpha})$, the marginals $X_i \sim beta(\alpha_1, A-\alpha_i)$.

Next, let us show that the sum of two Dirichlet marginal distributions is beta distributed: $Y_1 = X_1 + X_j$

The joint distribution of X_1, X_j:

$$f_{x_1,x_j}(x_1, x_j) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(A - \alpha_1 - \alpha_j)} x_1^{\alpha_1 - 1} x_j^{\alpha_j - 1}(1 - x_1 - x_j)^{A - \alpha_1 - \alpha_j - 1}$$

$$\alpha_0 = \sum_{j \neq \{1,2,j\}} \alpha_j$$

Applying the following transformation:

$$y_1 = x_1 + x_j$$

$$u = \frac{x_1}{x_1 + x_j}$$

$$x_1 = uy_1$$

$$x_j = y_1 - uy_1$$

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial u} \\ \frac{\partial x_j}{\partial y_1} & \frac{\partial x_j}{\partial u} \end{vmatrix} = \begin{vmatrix} u & y_1 \\ 1 - \mathrm{u} & -y_1 \end{vmatrix} = -uy_1 - y_1(1 - u) = -y_1$$

$$f_{y_1,u}(y_1, u) = f_{x_1,x_j}(y_1, u)|J|$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(A - \alpha_1 - \alpha_j)} (uy_1)^{\alpha_1 - 1}(y_1 - uy_1)^{\alpha_j - 1}(1 - uy_1 - y_1 + uy_1)^{A - \alpha_1 - \alpha_j - 1}|-y_1|$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(A - \alpha_1 - \alpha_j)} u^{\alpha_1 - 1} y_1^{\alpha_1 - 1} y_1^{\alpha_j - 1}(1 - u)^{\alpha_j - 1}(1 - y_1)^{A - \alpha_1 - \alpha_j - 1} y_1$$

Obtaining the distribution of Y_1 by integrating out *u*:

$$f_{y_1}(y_1) = \int_0^1 f_{y_1,u}(y_1,u)du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(A-\alpha_1-\alpha_j)} y_1^{\alpha_1+\alpha_j-1}(1-y_1)^{A-\alpha_1-\alpha_j-1}\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_j-1}du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(\alpha_0+\alpha_2)} y_1^{\alpha_1+\alpha_j-1}(1-y_1)^{\alpha_0+\alpha_2-1}\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_j-1}du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(\alpha_j)\Gamma(\alpha_0+\alpha_2)} y_1^{\alpha_1+\alpha_j-1}(1-y_1)^{\alpha_0+\alpha_2-1}\frac{\Gamma(\alpha_1)\Gamma(\alpha_j)}{\Gamma(\alpha_1+\alpha_j)}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_1+\alpha_j)\Gamma(\alpha_0+\alpha_2)} y_1^{\alpha_1+\alpha_j-1}(1-y_1)^{\alpha_0+\alpha_2-1}$$

This is the pdf of a beta distribution, and therefore: $Y_1 \sim beta(\alpha_1+\alpha_j, \alpha_0+\alpha_2)$

Analogously, sum of the two Dirichlet marginal distributions: $Y_2 = X_2 + X_j$

The joint distribution of X_2, X_j:

$$f_{x_2,x_j}(x_2, x_j) = \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(A - \alpha_2 - \alpha_j)} x_2^{\alpha_2 - 1} x_j^{\alpha_j - 1} (1 - x_2 - x_j)^{A - \alpha_2 - \alpha_j - 1}$$

$$\alpha_0 = \sum_{j \neq \{1,2,j\}} \alpha_j$$

Applying the following transformation:

$$y_2 = x_2 + x_j$$

$$u = \frac{x_2}{x_2 + x_j}$$

$$x_2 = uy_2$$

$$x_j = y_2 - uy_2$$

$$J = \begin{vmatrix} \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial u} \\ \frac{\partial x_j}{\partial y_2} & \frac{\partial x_j}{\partial u} \end{vmatrix} = \begin{vmatrix} u & y_2 \\ 1 - \mathrm{u} & -y_2 \end{vmatrix} = -uy_2 - y_2(1 - u) = -y_2$$

$f_{y_2,u}(y_2, u) = f_{x_2,x_j}(y_2, u)|J|$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(A - \alpha_2 - \alpha_j)} (uy_2)^{\alpha_2 - 1} (y_2 - uy_2)^{\alpha_j - 1} (1 - uy_2 - y_2 + uy_2)^{A - \alpha_2 - \alpha_j - 1} |-y_2|$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(A - \alpha_2 - \alpha_j)} u^{\alpha_2 - 1} y_2^{\alpha_2 - 1} y_2^{\alpha_j - 1} (1 - u)^{\alpha_j - 1} (1 - y_2)^{A - \alpha_2 - \alpha_j - 1} y_2$$

Obtaining the distribution of Y_2 by integrating out u:

$$f_{y_2}(y_2) = \int_0^1 f_{y_2,u}(y_2, u)du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(A-\alpha_2-\alpha_j)} y_2^{\alpha_2+\alpha_j-1}(1-y_2)^{A-\alpha_2-\alpha_j-1}\int_0^1 u^{\alpha_2-1}(1-u)^{\alpha_j-1}du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(\alpha_0+\alpha_1)} y_2^{\alpha_2+\alpha_j-1}(1-y_2)^{\alpha_0+\alpha_1-1}\int_0^1 u^{\alpha_1-1}(1-u)^{\alpha_j-1}du$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2)\Gamma(\alpha_j)\Gamma(\alpha_0+\alpha_1)} y_2^{\alpha_2+\alpha_j-1}(1-y_2)^{\alpha_0+\alpha_1-1}\frac{\Gamma(\alpha_2)\Gamma(\alpha_j)}{\Gamma(\alpha_2+\alpha_j)}$$

$$= \frac{\Gamma(A)}{\Gamma(\alpha_2+\alpha_j)\Gamma(\alpha_0+\alpha_1)} y_1^{\alpha_2+\alpha_j-1}(1-y_2)^{\alpha_0+\alpha_1-1}$$

This is the pdf of a beta distribution, and therefore: $Y_1 \sim beta(\alpha_2+\alpha_j, \alpha_0+\alpha_1)$

In the next section we derive an expression for correlation:

We know:

$$Y_1 \sim Beta(\alpha_1 + \alpha_j, \alpha_0 + \alpha_2)$$

$$Y_2 \sim Beta(\alpha_2 + \alpha_j, \alpha_0 + \alpha_1)$$

$$E[Y_1] = \frac{\alpha_1 + \alpha_j}{\alpha_1 + \alpha_j + \alpha_0 + \alpha_2}$$

$$E[Y_2] = \frac{\alpha_2 + \alpha_j}{\alpha_2 + \alpha_j + \alpha_0 + \alpha_1}$$

$$var(Y_1) = \frac{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)}{(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2)^2(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2 + 1)}$$

$$var(Y_2) = \frac{(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}{(\alpha_2 + \alpha_j + \alpha_0 + \alpha_1)^2(\alpha_2 + \alpha_j + \alpha_0 + \alpha_1 + 1)}$$

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{sd(Y_1)sd(Y_2)}$$

Let's find an expression for the covariance:

$$Cov(Y_1, Y_2) = E[(Y_1 - E[Y_1])(Y_2 - E[Y_2])]$$

$$= E[Y_1Y_2 - Y_1E[Y_2] - Y_2E[Y_1] + E[Y_1]E[Y_2]]$$

$$= E[Y_1Y_2] - E[Y_1]E[Y_2] - E[Y_2]E[Y_1] + E[Y_1]E[Y_2]$$

$$= E[Y_1Y_2] - E[Y_1]E[Y_2]$$

$$= E[Y_1Y_2] - \frac{(\alpha_1 + \alpha_j)(\alpha_2 + \alpha_j)}{(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2)^2}$$

$$= E[Y_1Y_2] - \frac{(\alpha_1 + \alpha_j)(\alpha_2 + \alpha_j)}{A^2}$$

Let's find $E[Y_1Y_2]$:

$$E[Y_1Y_2] = E\big[(X_1 + X_j)(X_2 + X_j)\big]$$

$$= E[X_1X_2] + E\big[X_1X_j\big] + E\big[X_2X_j\big] + E\big[X_j^2\big]$$

Looking at the four terms separately:

$$E[X_1X_2] = E[X_1]E[X_2] + cov(X_1, X_2)$$

The covariance between X_1 and X_2 if $\vec{X} \sim Dirichlet(\vec{\alpha})$ is $cov(X_1, X_2) = \frac{-\alpha_1\alpha_2}{A^2(A+1)}$

$$E[X_1X_2] = \frac{\alpha_1}{\alpha_1 + A - \alpha_1}\frac{\alpha_2}{\alpha_2 + A - \alpha_2} + \frac{-\alpha_1\alpha_2}{A^2(A+1)}$$

$$= \frac{\alpha_1\alpha_2(A-1)}{A^2(A-1)} + \frac{-\alpha_1\alpha_2}{A^2(A+1)}$$

$$= \frac{\alpha_1\alpha_2 A + \alpha_1\alpha_2 - \alpha_1\alpha_2}{A^2(A+1)}$$

$$= \frac{\alpha_1\alpha_2}{A(A+1)}$$

Analogously:

$$E\big[X_1X_j\big] = \frac{\alpha_1\alpha_j}{A(A+1)}$$

$$E\big[X_2X_j\big] = \frac{\alpha_2\alpha_j}{A(A+1)}$$

$$E\big[X_j^2\big] = var\big(X_j\big) + E\big[X_j\big]^2 = \frac{\alpha_j\big(A - \alpha_j\big)}{\big(\alpha_j + A - \alpha_j\big)^2\big(\alpha_j + A - \alpha_j + 1\big)} + \left(\frac{\alpha_j}{\alpha_j + A - \alpha_j}\right)^2$$

$$= \frac{\alpha_j\big(A - \alpha_j\big)}{A^2(A+1)} + \frac{\alpha_j^2}{A^2}$$

$$= \frac{\alpha_j A - \alpha_j^2 + \alpha_j^2 A + \alpha_j^2}{A^2(A+1)} = \frac{\alpha_j A + \alpha_j^2 A}{A^2(A+1)} = \frac{\alpha_j\big(\alpha_j + 1\big)}{A(A+1)}$$

Plugging in the four terms:

$$E[Y_1Y_2] = \frac{\alpha_1\alpha_2}{A^2} + \frac{\alpha_1\alpha_j}{A^2} + \frac{\alpha_2\alpha_j}{A^2} + \frac{\alpha_j(\alpha_j+1)}{A(A+1)}$$

$$Cov(Y_1, Y_2) = \frac{\alpha_1\alpha_2}{A(A+1)} + \frac{\alpha_1\alpha_j}{A(A+1)} + \frac{\alpha_2\alpha_j}{A(A+1)} + \frac{\alpha_j(\alpha_j+1)}{A(A+1)} - \frac{(\alpha_1+\alpha_j)(\alpha_2+\alpha_j)}{A^2}$$

$$= \frac{\alpha_1\alpha_2A + \alpha_1\alpha_jA + \alpha_2\alpha_jA + \alpha_j(\alpha_j+1)A}{A^2(A+1)} - \frac{(\alpha_1+\alpha_j)(\alpha_2+\alpha_j)(A+1)}{A^2(A+1)}$$

$$= \frac{\alpha_1\alpha_2A + \alpha_1\alpha_jA + \alpha_2\alpha_jA + \alpha_j^2A + \alpha_jA}{A^2(A+1)}$$

$$+ \frac{-\alpha_1\alpha_2A - \alpha_1\alpha_jA - \alpha_j\alpha_2A - \alpha_j^2A - \alpha_1\alpha_2 - \alpha_1\alpha_j - \alpha_j\alpha_2 - \alpha_j^2}{A^2(A+1)}$$

$$= \frac{\alpha_jA - \alpha_1\alpha_2 - \alpha_1\alpha_j - \alpha_j\alpha_2 - \alpha_j^2}{A^2(A+1)}$$

$$= \frac{\alpha_j(A - \alpha_1 - \alpha_2 - \alpha_j) - \alpha_1\alpha_2}{A^2(A+1)}$$

$$= \frac{\alpha_0\alpha_j - \alpha_1\alpha_2}{A^2(A+1)}$$

Now let's work through the denominator:

$$sd(Y_1)sd(Y_2) = \sqrt{\frac{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)}{(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2)^2(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2 + 1)}}$$

$$\cdot \sqrt{\frac{(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}{(\alpha_2 + \alpha_j + \alpha_0 + \alpha_1)^2(\alpha_2 + \alpha_j + \alpha_0 + \alpha_1 + 1)}}$$

$$= \sqrt{\frac{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}{(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2)^4(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2 + 1)^2}}$$

$$= \frac{\sqrt{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}}{(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2)^2(\alpha_1 + \alpha_j + \alpha_0 + \alpha_2 + 1)}$$

$$= \frac{\sqrt{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}}{A^2(A+1)}$$

Combining the nominator and denominator:

$$Corr(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{sd(Y_1)sd(Y_2)}$$

$$= \frac{\frac{\alpha_0\alpha_j - \alpha_1\alpha_2}{A^2(A+1)}}{\frac{\sqrt{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}}{A^2(A+1)}}$$

$$= \frac{-\alpha_1\alpha_2 + \alpha_0\alpha_j}{\sqrt{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}}$$

This is an expression for the correlation between Y_1 and Y_2.

Next, we will describe a sampling procedure:

As described above the correlation is given by:

$$r = Corr(Y_1, Y_2) = \frac{-\alpha_1\alpha_2 + \alpha_0\alpha_j}{\sqrt{(\alpha_1 + \alpha_j)(\alpha_0 + \alpha_2)(\alpha_2 + \alpha_j)(\alpha_0 + \alpha_1)}}$$

$$c_1 = \alpha_1 + \alpha_j$$

$$c_2 = \alpha_0 + \alpha_2$$

$$c_3 = \alpha_2 + \alpha_j$$

$$c_4 = \alpha_0 + \alpha_1$$

Rearranging:

$$\alpha_1 = c_1 - \alpha_j$$

$$\alpha_j = c_3 - \alpha_2$$

$$\alpha_2 = c_2 - \alpha_0$$

$$\alpha_0 = c_4 - \alpha_1$$

$$\alpha_1 = c_1 - \alpha_j$$

$$\alpha_1 = c_1 - c_3 + \alpha_2$$

$$\alpha_1 = c_1 - c_3 + c_2 - \alpha_0$$

$$\alpha_1 = c_1 - c_3 + c_2 - c_4 + \alpha_1$$

$$c_4 = c_1 + c_2 - c_3$$

Since $c_1, \ldots c_4$ are shape parameter of a beta distribution, $c_i > 0$

$$c_4 > 0$$

$$c_1 + c_2 > c_3$$

Recall the correlation r:

$$r = \frac{-\alpha_1\alpha_2 + \alpha_0\alpha_j}{\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)}}$$

$$\alpha_0\alpha_j - \alpha_1\alpha_2 = r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)}$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - \alpha_1)\alpha_j - (c_1 - \alpha_j)(c_2 - \alpha_0)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - c_1 + \alpha_j)\alpha_j - (c_1 - c_3 + \alpha_2)(c_2 - c_4 + \alpha_1)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - c_1 + \alpha_j)\alpha_j - (c_1 - c_3 + c_2 - \alpha_0)(c_2 - c_4 + c_1 - \alpha_j)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - c_1 + \alpha_j)\alpha_j - (c_1 - c_3 + c_2 - c_4 + \alpha_1)(c_2 - c_4 + c_1 - \alpha_j)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - c_1 + \alpha_j)\alpha_j - (c_1 - c_3 + c_2 - c_4 + c_1 - \alpha_j)(c_2 - c_4 + c_1 - \alpha_j)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = (c_4 - c_1 + \alpha_j)\alpha_j - (c_1 - \alpha_j)(c_3 - \alpha_j)$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} = \alpha_j c_4 - \alpha_j c_1 + \alpha_j^2 - c_1c_3 + c_1\alpha_j + c_3\alpha_j - \alpha^2$$

$$r\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} + c_1c_3 = \alpha_j(c_4 + c_3)$$

$$\alpha_j = r\frac{\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} + c_1c_3}{c_1 + c_2}$$

So, the α's can be expressed as a function of *c*'s and *r*:

$$\alpha_j = r\frac{\sqrt{c_1c_2c_3(c_1 + c_2 - c_3)} + c_1c_3}{c_1 + c_2}$$

$$\alpha_1 = c_1 - \alpha_j$$

$$\alpha_2 = c_3 - \alpha_j$$

$$\alpha_0 = c_4 - c_1 + \alpha_j = c_2 - c_3 + \alpha_j$$

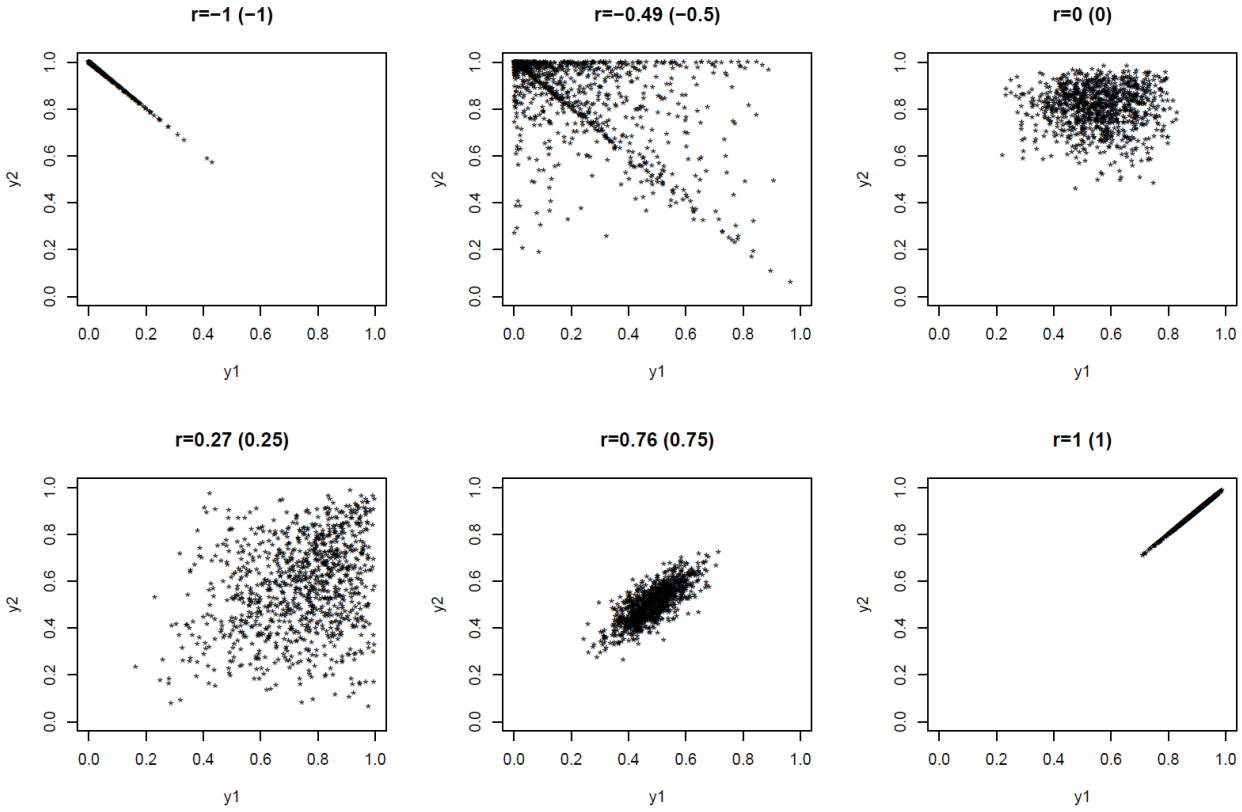
Procedure for sampling:

1. define values for $c_1, c_2, c_3 (, c_4)$ and r
2. make sure that $c_3 < c_1 + c_2$ and $c_i > 0$
3. calculate $\alpha_1, \alpha_2, \alpha_j, \alpha_0$
4. make sure that $\alpha_i > 0$
5. draw n samples from: $\vec{X} \sim dir(\alpha_1, \alpha_2, \alpha_j, \alpha_0)$
6. calculate sums of X_1, X_2, X_j for correlated beta values:
    a. $Y_1 = X_1 + X_j$
    b. $Y_2 = X_2 + X_j$

The following code will create correlated observations, that follow a beta distribution with the correlation $r = \{-1.00, -0.50, 0.00, 0.25, 0.75, 1.00\}$:

```
library("MCMCpack")
for(r in c(-1, -0.5, 0, 0.25, 0.75, 1)){
  aj = a1 = a2 = a0 = -1
  while(a1 <= 0 | a2 <= 0 | aj <= 0 | a0 <= 0){ #ai needs to be positive
    c1 = rgamma(1,1,0.1)
    c2 = rgamma(1,1,0.1)
    if(r == 1){
      c3 = c1
    } else if(r == -1){
      c3 = c2
    }else{
      c3 = c1+c2
      while(c3>=c1+c2) c3 = rgamma(1,1,0.1)
    }
    aj = (r * sqrt(c1*c2*c3*(c1+c2-c3)) + c1*c3) / (c1+c2)
    a1 = c1 - aj
    a2 = c3 - aj
    a0 = c2 - c3 + aj
  }
  myDir = rdirichlet(1000, c(a1, a2, aj, a0))
  y1 = myDir[,1] + myDir[,3]
  y2 = myDir[,2] + myDir[,3]
  plot(y1,y2, pch="*", xlim = c(0,1), ylim = c(0,1), main =paste("r=", round(cor(y1,y2),2), " (", r,")", sep = ""))
}
```

This code proves the following scatterplots illustrating the specified correlations. Each subfigure's title indicates the calculated correlation from the simulated specified and in parenthesis specified correlation:


r=−1 (−1)
r=−0.49 (−0.5)
r=0 (0)
r=0.27 (0.25)
r=0.76 (0.75)
r=1 (1)
y1
y2
0.0
0.2
0.4
0.6
0.8
1.0