Differential Network Analysis and Validation Strategies for High-dimensional Oncological Genetic Data

Dissertation by

Miriam Lohr

submitted to the Department of Statistics,

Technische Universität Dortmund, Germany

in fulfillment of the requirements for

the degree Doktor der Naturwissenschaft

submitted November 2014

Date of oral examination: February 6, 2015

Primary supervisor: Prof. Dr. Jörg Rahnenführer

Secondary supervisor: Dr. Uwe Ligges

Contents

1 Introduction **1**

2 Biological background and datasets **8**
2.1 Cancer 8
2.2 Datasets 13
2.2.1 Breast cancer datasets 14
2.2.2 Non-small cell lung cancer datasets 15

3 Validation approaches for high-dimensional genetic data **19**
3.1 Methods 22
3.1.1 Multiple testing 22
3.1.2 Cox proportional hazards model 25
3.1.3 Meta-analysis 28
3.2 Sequential validation strategy 33
3.2.1 Stepwise validation on breast cancer datasets 34
3.2.2 Performance check via simulation studies 36
3.2.3 Application to non-small cell lung cancer datasets 41
3.3 Two-step meta-analysis approach 45
3.4 Summary and comparison 47

4 Differential gene expression networks **55**
4.1 Methods for differential network analysis 57
4.1.1 The link from probability theory to graph theory 57
4.1.2 Partial correlations 62
4.1.3 Shrinkage estimation of the covariance matrix 67
4.1.4 The local False Discovery Rate 71
4.1.5 Measures for the comparison of two undirected graphs 72
4.1.6 Permutation tests for the quantification of differential networks . 78
4.2 Detection of differential genetic networks 80
4.2.1 Predefined networks from literature 80
4.2.2 DiNGS - Gene selection for differential networks 84
4.3 A simulation study for the detection of differential neworks 92

4.3.1 Design of data 92
4.3.2 Properties of tests for the quantification of differential networks . 95
4.4 Summary 110

5 Discussion and conclusions **115**

References **119**

Acknowledgements **138**

Declaration **139**

Appendix **140**
Additional tables 140
Additional figures 156

1 Introduction

Cancer — one of three women and one of two men receive this diagnosis during their lifetime (US National Cancer Institute, 2013a, 2013b, 2013c). Although many new therapies have been developed in the past decades, the lifetime risk of dying from cancer is still about 20% (22.94% and 19.34% in 2012 in the US for males and females, respectively).
Tumorigenesis is caused by an imbalance of proliferation and programmed cell death. Genes that regulate these mechanisms are altered (Croce, 2008) and can be divided into two categories. Oncogenes are responsible for proliferation and cell growth, and tumor supressor genes promote cell death (Wilbur, 2009). Recent research aims to get detailed insights into cancer biology to target mechanisms responsible for tumor development and progression.
Changes in gene expression is one of several indications for genetic alterations in cancer. Knudson (2001) discovered that typically many genes are needed to change a normal cell into a tumor cell with uncontrolled growth. Microarray technology, and most recently RNA-seq, is used to measure the expression of thousands of genes simultaneously. Hundreds of cancer-related gene expression datasets are publicly available in databases like Gene Expression Omnibus (Edgar et al., 2002).

Since modulation of gene expression is caused by either chromatin domains, transcription, post-transcriptional modification, RNA transport, translation or mRNA degradation (Gilbert, 2003), thousands of simultaneously measured gene expressions are promissing to reveal mechanisms changing under certain conditions.
With the first microarray experiments, researchers have been started to look for differentially expressed genes between disease and control samples, different stages of a disease

or different tissues. For example, Ismail et al. (2000) identified 160 and 95 genes up-regulated in normal human ovarian surface epithelium and ovarian tumors, respecively. Welsh et al. (2001) found several differentially expressed genes between prostate tumors and normal prostate tissue and proposed the secreted macrophage inhibitory cytokine (MIC-1) as diagnostic marker.

In this thesis, our goal is to improve statistical methods for high-dimensional data to gain deeper insights into cancer biology by analyzing gene expression data. We focus on two topics. First, the large number of available gene expression datasets is used to validate differentially expressed genes or biomarkers in cancers. Second, genes are not considered alone but in interaction networks that might change during disease progression or between different disease stages. We detect differential interaction networks from gene expression data by testing gene sets derived with and without biological prior knowledge.

Despite the knowledge of multi-gene involvement in tumorigenesis, many publications describe single genes as predictive or prognostic markers. Trock et al. (1997) found that MDR1/gp170 expression in breast cancer tumors is associated with poor response to chemotherapy and Yamabuki et al. (2007) identified Dikkopf-1 as a novel serologic and prognostic biomarker for lung and esophageal carcinomas.
The growing number of available data offers the opportunity to validate or disprove findings. A common approach to evaluate the prognostic or predictive impact of a gene is meta-analysis (Whitehead, 2002). E.g. Griffith et al. (2006) identified important diagostic biomarkers of thyroid cancer via meta-analysis. Mehra (2005) discovered GATA3 as prognostic marker in breast cancer by global gene expression meta-analysis.
Common meta-analysis treats all considered datasets equally except for different weighting. But often findings are discovered on one dataset and shall be validated on other homogeneous cohorts. Moreover, in all genome-wide expression analyses prognostic or predictive genes may be misleadingly found due to the high number of performed tests. To control the number of false positives a correction for multiple testing is required and commonly performed.
Bonferroní (Hochberg and Tamhane, 1987) published a simple procedure to control the global number of false positive results. Holm (1979) proposed a less conservative method and shows that it still controls the global type one error. A less strict approach was pro-

posed by Benjamini and Hochberg (1995) by controlling the proportion of false positives among all significant findings. The latter is commonly used for correction for multiple testing in high-dimensional genetic data, particularly because expression profiles of genes are not independent.
Taking the validation idea and the adjustment for multiple testing into consideration, Miller et al. (2001) proposed "a two-stage design in which significance testing applied to explorytory data is used to guide a second round of hypothesis-testing experiment conducted in a separate set of experimental studies". Victor and Hommel (2007) combine an adaptive design with the control of the false discovery rate and argue for a generalized definition for a global p-value. Zehetmayer and Posch (2012) proposed an integrative approach that is based on the pooled data from both stages in a two stage approach controlling the FDR.

However, the idea behind the listed approaches is to reduce experimental costs, but not to validate previous findings. In this thesis, we propose two new approaches to validate biomarkers derived from high-dimensional data. The first strategy combines an exploratory screening for markers with a common meta-analysis of validation datasets. The second approach is based on sequential validation of considered datasets. By successively reducing the number of genes through the validation steps less adjustment for multiple testing is required. Both approaches are already puplished by Lohr et al. (2012) and Botling et al. (2013), respectively.

In the past years biological network interference has become a major research topic, because researchers realized that — especially cancer — biology is more complex and cannot be assessed by analyzing the expression of single genes.
Many methods for network inference have been proposed in the past decades. Butte et al. (2000) published the concept of relevance networks that use the Pearson correlation and a fixed threshold to determine which edges are present in a graph. A more sophisticated approach for graph estimation are Bayesian Networks (Pearl, 2000). Sampling from the posterior distribution of the graph given the observed data via Markov Chain Monte Carlo (MCMC) simulations (Grzegorczyk and Husmeier, 2008), the posterior probability for each edge is determined by the average of simulated samples. Although Friedman et al. (2000) applied this method to expression data, inference of Bayesian

Networks is computational expensive due to the MCMC simulations and therefore not suitable for high-dimensional data.
An alternative method for network inference is the Covariance Selection or Graphical Gaussian Models based on an idea from Dempster (1972). Whittaker (1990) developed the theory of Graphical Gaussian Models assuming a multivariate normal distribution of the data. Zero entries of the inverse covariance matrix, i.e. the precision matrix, and therefore in the matrix of partial correlations are interpreted as absent edges in a graph, while none-zero entries in the matrix of partial correlations denote present edges. Partial correlation denotes the correlation of two variables if the influence of other variables is removed. For the calculation of the partial correlation matrix, the inverse of the covariance matrix is required. The covariance matrix can only be inverted if it has full rank, i.e. more observations than variables must be considered. However, gene expression is often measured for only few samples.

Many researchers adopt the idea of the Graphical Gaussian Models and developed methods that allow sparse estimation of the covariance or rather precision matrix for genetic data.
Schäfer and Strimmer (2005a) proposed a linear shrinkage approach for the estimation of the covariance matrix. By combining the unconstrained estimation of the covariance matrix with a constrained estimator — a diagonal matrix with the variances of genes — the resulting estimation of the covariance matrix will be positive definit and therefore invertible.
Friedmann et al. (2008) applied a Lasso penalty to the covariance matrix. Their proposed algorithm fits a modified Lasso regression iteratively for each variable. The idea for the algorithm was adopted from Meinshausen and Bühlmann (2005). They combined neighborhood selection that estimates the conditional independence restrictions separately for each node with the Lasso as an alternative to standard covariance selection for sparse high-dimensional graphs.
Another approach for the regularized estimation of the covariance and thereof partial correlation matrix was proposed by Tenenhaus et al. (2010). They applied a Partial Least Squares Regression to assess the strength of independence of any two genes in a small-sample-size and high-dimensional network setting.
A general framework for combining regularized regression methods with the estimation

of Graphical Gaussian models was introduced by Krämer et al. (2009). They suggested to use various existing methods like Partial Least Squares Regression as well as two new approaches based on ridge regression and two-stage adaptive lasso, comparing sparse and non-sparse methods for gene-association estimation. Extensive simulations and comparisons resulted in the conclusion that the shrinkage approach proposed by Schäfer and Strimmer (2005a) is more stable than regression based methods like the one from Meinshausen and Bühlmann (2005).

Recently many methods for graph comparisons in the context of microarray data have been developed. De la Fuente (2010) resumes the recent ideas published from differential expression of single genes to differential coexpression and further to different (interaction) networking. Choi et al. (2005) compared tumor and normal tissue by estimating Relevence Networks for each phenotype. Assuming the same number of edges to be present in both graphs, edges that are exclusive in one of the two graphs are called subtype-specific links.
A method to detect changes between multiple ordered groups, e.g. time series, was proposed by Gillis and Pavlidis (2009). Differential co-expression between multiple groups is assessed by a measure based on Haar-wavelets (Haar, 1909).
Jacob et al. (2012) combined both steps — first testing individual genes, then testing gene sets for enrichment of differentially expressed genes — in a single procedure. Their method takes the network topology, e.g. from KEGG pathways (Kanehisa and Goto, 2000), into account to gain more power.

These methods for graph comparison depend on differential co-expression of gene networks. An alternative approach is to quantify a change in the interaction structure of gene expression.
Gill et al. (2010) proposed a framework for the detection of differential connectivity. They use a connectivity score to test whether the overall modular structure of two graphs, the connetivity of a specific set of "interesting genes", or the connectivity of a single gene between two networks is different. Therefore, a permutation test using the mean distance of partial correlations derived by the shrinkage approach of Schäfer and Strimmer (2005a) is applied.
A method for Indirect Comparisons of Interaction Graphs was proposed by Mansmann

et al. (2010). A hierarchical top-down testing approach using resampling technique is applied beginning with the global null-hypothesis "no node in the network shows a different interaction".

To assess whether the interaction of genes is different under two conditions, it is not feasible to consider the entire collection of measured genes. Hence, strategies for hypothesis generation of differential networks are required.
Kostka and Spang (2004) detected sets of differentially co-expressed genes under two conditions by calculating a score based on an ANOVA model for differential co-expression and application of an algorithm that finds high scoring gene sets.

Gambardella et al. (2013) developed a procedure named DINA (DIfferential Network Analysis) that is able to identify a set of genes, whose co-regulation is condition-specific. DINA starts with a set of genes, e.g. a KEGG pathway, and a set of networks, for example derived by Spearman Correlation analysis. A co-regulation probability is calculated and its variability across networks is assessed based on permutation testing of an entropy which describes the uncertainty associated with a random variable, i.e. the genes have a high co-regulation probability only in one network, i.e. the pathway activity is condition-specific, the entropy will be low.
Another approach is to test gene sets defined by biological prior knowledge. For example Jacob et al. (2012) proposed to apply their testing procedure to all KEGG pathways.

However, due to the large number of genes and constrains of used methods, the detection of differential interaction networks remains challenging. In this thesis, we propose a new algorithm for Differential Networks Gene Selection(DiNGS). Starting with a suitable pair of genes a forward selection is performed on a criterion ensuring differential co-regulation between two groups.
Adopting the idea of differential expression to our aim to find differential interaction networks, we perform a Gene Set Enrichment Analysis in Gene Ontology (GO) groups (Edgar et al., 2002) of genes known to have impact on breast cancer prognosis. Enriched gene sets are afterwards tested for differences in their corresponding interaction networks of breast cancer patients with and without metastasis by several permutation

tests. These tests use test statistics based on partial or orinary correlations including the test proposed by Gill et al. (2010). The properties of the permutation tests are explored in an extensive simulation study.
Most of the permutation tests are already published in Lohr et al. (2010).

This thesis is organized as follows: In Chapter 2 we provide the biological background of cancer and describe the datasets considered in this thesis. Chapter 3 introduces the validation approaches for high-dimensional gene expression data. Within this chapter, Section 3.1 presents methods for meta-analysis and validation. The sequential validation procedure and the two-step meta-analysis approach are introduced in Sections 3.2 and 3.3, respectively. Section 3.4 summarizes the results of both methods applied to non-small cell lung cancer and compares them with an ordinary meta-analysis.
In Chapter 4, concepts for the detection of differential networks are introduced. The corresponding methods are described in Section 4.1. Section 4.2 presents explicit ways to discover differential interaction networks. A simulation study to explore the properties of the permutation tests introduced in Section 4.1 is presented in Section 4.3. Results of this chapter are summerized in Section 4.3.
Chapter 5 discusses methods and findings from both topics — the validation approaches for high-dimensional gene expression data and the detection of differential gene interaction sub-networks — and gives an outlook to possible extensions and provides concluding remarks of this work.

2 Biological background and datasets

2.1 Cancer

In Germany cancer is the most frequent cause of death after cardiovascular diseases (Kaatsch et al., 2012). Worldwide, more than 12 million people are newly diagnosed with cancer per year (Jemal et al., 2011). In fact, cancer comprises more than 200 different diseases (Schulz, 2005). Though all cancers share the same elementary features in matters of malignancy, it shows a large range of diversity, which requires different therapeutic strategies.

The DNA contains the information that is required for an organism to develop its mass and shape, and the information about every protein that is needed for biological processes. The central dogma of molecular biology describes the flow of information in a biological organism consisting of replication, transcription, and translation. In the replication the DNA sequence is copied to transfer it from a mother to a daughter cell. In the transcription, DNA is rewritten into mRNA. Afterwards the information can be translated through the mRNA to a protein specified by the DNA sequence (Alberts et al., 2007). Although the organism has developed many controls to avoid errors in replication, errors sometimes do occur, which might lead to erroneous incorporation into the newly synthesised DNA strand due to mutated nucleotides and may cause a disequilibrium between cell growth and apoptosis. Though the diversity of cancers is high, some cancer types affect the people more than others. This thesis is focused on breast and non-small cell lung cancer that play a major role in terms of world-wide incidence and mortality.

Breast cancer

Carcinomas of the breast are the most frequent cancer types in women by far with a age-standardized incidence of 123.1 per 100 000 female inhabitants in 2008 in Germany (Kaatsch et al., 2012). Breast cancer caused 458 503 deaths worldwide which is 13% of all cancer-related deaths in women. Although women are 100 times more frequently affected than men, men may also be affected and they tend to have poorer outcomes due to delays in diagnosis (American Cancer Society, 2013). The risk for breast cancer increases with age, i.e. postmenopausal, but also younger women are affected and often with poorer prognosis due to hereditary predispositions. Other known risk factors are long (life-)time exposure of estrogens, ionizing radiation, cigarette smoking, alcohol, and a high-fat diet (Schulz, 2005), where several factors might interact, also synergistically. Breast carcinomas are classified by several aspects. The main aspects, histopathology, stage, grade, and receptor status, are considered for treatment selection and conclusions may be drawn for prognosis.
First, breast cancer is classified by its histological appearance. About three quarters of all breast cancers are invasive ductal carcinomas (55%), ductal carcinomas in situ (13%), and invasive lobular carcinomas (5%) (Eheman et al., 2009). The World Health Organization (WHO) recommends further subdivision of breast cancers according to pathological type (WHO, 2003).
The determination of amount and location of the cancer in the organism or body is called staging. We know two different kinds of staging, the clinical staging that is obtained by mammography, x-rays and CT scans before surgery, and the pathological staging obtained by surgery which is more accurate (American Joint Committee on Cancer, 2010). The TNM classification system, a commom used scheme for several cancer types, that is based on the size of the tumor (T), the invasion of surrounding organs and lymph nodes (N), and the presence of distant metastasis (M) (Sobin and Compton, 2010). The classification is explained below in the following section about non-small lung cancer, because staging information is available for the lung cancer datasets but not for all breast cancer data used in this thesis.

However, the grading is used for breast cancer classification in this thesis. It depends on the microscopic appearance of the breast cancer cells compared to normal breast cells.

Grading classifies the tissue in well differentiated (low grade), moderately differentiated (intermediate grade), and poorly differentiated (high grade) tumors. In the following, the Nottingham (also called Elston-Ellis) scale system (Elston and Ellis, 1991) is used as modification of the Scarff-Bloom-Richardson grading system (Bloom and Richardson, 1957), which grades breast carcinomas by summing scores for tubule formation, nuclear pleomorphism, and mitotic count. The score ranges from I to III, where I stands for well differentiated, while a poor or undifferentiated tumor is given a higher score of III.
Another classification criterion is the status of hormon receptors like estrogen receptors (ER), progesterone receptors (PR) and HER2/ERBB2. The presence of receptors is often identified by immunohistological analysis. Estrogen in combination with its receptor is a key regulator of growth in a normal breast (Schulz, 2005). ER positive (ER+) tumors depend on estrogen for their growth and therefore may be treated with drugs to reduce the effect of estrogen, e.g. Tamoxifen. HER2 (human epidermal growth factor receptor 2) is a protein encoded by the ERBB2 gene (Coussens et al., 1985). It stimulates cell proliferations and inhibits apoptosis. Patients with overexpression of ERBB2 ("HER2 positiv") have a poorer prognosis, but treatment by a monoclonal antibody "Trastuzumab" that binds at HER2 is indicated. PR is a protein inside cells that is activated by progesterone (Law et al., 1987) and has functions in maintaining pregnancy, in estrous and menstrual cycles. Hence, the combination ER+/PR+/ERBB2- indicates a comparatively good prognosis.

Non-small cell lung cancer

Lung cancer is the leading cause of cancer-related death worldwide (Ferlay et al., 2010). More than 1.6 million people were newly diagnosed and about 1.38 million died due to lung cancer in 2008. Among men, lung cancer is more frequently diagnosed than among women, but the incidence in women has considerably increased over the last decades and has just recently begun to stabilize (Jemal et al., 2004). This fact can be explained by the amplified tobacco usage in women (Lum et al., 2008). Tobacco smoke is the most prominent risk factor that causes about $80 - 90\%$ of all lung carcinomas (Horn et al., 2012). Other known risk factor are genetic factors, radon gas, asbestos, and air pollution (Alberg and Samet, 2010; O'Reilly et al., 2007) including second-hand smoke (Carmona,

2006). Most patients diagnosed with lung cancer are older than 60 years (DKFZ, 2013). Compared to smoking-related lung cancer, carcinomas in non-smokers occur more often in women and are more often classified as so called adenocarcinomas (Subramanian and Govindan, 2007). In gerenal, lung cancer can be divided into small cell lung cancers (SCLC) that account for approximately 15% of all lung cancers and non-small cell lung cancers (NSCLC) (Travis, 2011). SCLC is assumed to have its origin in neuroendocrine cells of the lung (Rosti et al., 2006). Patients diagnosed with SCLC show a promising response to chemotherapy at first, but often develop a therapy resistance followed by metastatic disease within five years. In this thesis we focus on datasets consisting of NSCLC.

By cellular morphology, three main subgroups of histological substypes are defined which are the above mentioned adenocarcinomas, squamous cell and large cell carcinomas that make up 40%, 21%, and 14% of all diagnosed lung cancers, respectively. A microscopic examination of the stained tissue is a standard procedure after surgery. The classification of histological subtype is essential for the choice of therapy (Langer et al., 2010). The glandular structure is characteristic for *adenocarcinomas* as well as the production of mucin (Cooper et al., 2011). It is the most common type in non-smokers as mentioned above and in men younger than 50 years as well as in women of all ages. Adenocarcinomas are also associated with KRAS or EGFR mutations (Sequist et al., 2007). Keratinisation and intercellular bridges are typically seen in squamous cell carcinomas, while large cell carcinomas are undifferentiated with no sign of glandular or squamous differentiation. Other minor histological subtypes of NSCLC are adenosquamous carcinomas, sacromatoid carcinomas and typical/atypical carcinoids (Travis, 2011).

The staging of a tumor as already mentioned in the breast cancer section above is defined by the TNM system (Goldstraw et al., 2007; Sobin and Compton, 2010). Here, the stage of cancer is evaluated by the tumor size (T1-T4), the invasion of lymph nodes and organ structures (N0-N4) and the presence of distant metastasis (M0-M1). Detailed information on the classification can be found in Table 20 in the appendix. By combining the classification of T, N, and M a stage in range of I to IV is assigned (cf. Table 1). The stages are further subdivided in "a" and "b", except for stage IV. If distant metastases are present, the size of the tumor or whether lymph nodes are affected does not matter.

Stage	TNM subset	Stage	TNM subset
IA	T1a/T1b N0 M0	IB	T2a N0 M0
IIA	T1a/T1b N1 M0	IIB	T2b N1 M0
	T2a N1 M0		T3 N0 M0
	T2a N0 M0		
IIIA	T1/T2 N2 M0	IIIB	T4 N2 M0
	T3 N1/N2 M0		any T N3 M0
	T4 N0/N1 M0		
IV	any T any N M1a/M1b		

Table 1: Tumor stage based on TNM (7. edition), reproduced from Tsim et al. (2010).

All metastatic cancers are classified as stage IV. The prognosis and treatment decision depend on the tumor stage.

Frequently, NSCLC is diagnosed at late stage, because of the absence of lung-cancer specific symptoms. First symptoms may be respiratory ones, like coughing, hoarseness, or chest pain. Further symptoms such as weight loss, headache, and fatique might indicate late stage cancer with presence of distant metastases. In stage I and II surgical resection is recommended with the chance of total remission, hence $30 - 40\%$ of stage I patients experience a tumor relapse (Spiro et al., 2007). Stage II patients are treated with adjuvant chemotherapy, but for stage I it is controversially discussed (Scott et al., 2007). Stage III is quite heterogeneous. Patients diagnosed with stage IIIa have a considerably higher 5-year survival rate than stage IIIb patients with 23% and 7%, respectively. Often patients with stage IIIa are surgically treated followed by adjuvant chemotherapy, while stage IIIb tumors are inoperable and recommended to be treated with a combination of radio- and chemotherapy (Jett et al., 2007). NSCLC with metastatic disease, i.e. stage IV, is considered to be incurable and patients are treated with combinated therapies to improve life quality and gain some more months or even years (Socinski et al., 2007).

2.2 Datasets

A *biomarker* is an indicator of underlying biology (Biomarker Definitions Working Group, 2001). We distinguish between predictive biomarkers that provide information about e.g. response to treatment and prognostic biomarkers which predict the outcome of individual patients e.g. in terms of overall survival times or relapse-free survial. For the breast and non-small cell lung cancer datasets that will be described below, different event-free survival times are available. The discovery and validation of prognostic biomarkers will be part of this thesis. The challenge is due to the data structure.
All considered datasets comprise gene expression data of thousands of genes measured on Affymetrix microarrays. In general, *microarray technology* bases on hybridisation of complementary DNA or RNA nucleotide strands located on a chip to fixed DNA molecules so that each spot represents a specific gene or transcript for thousands of genes in parallel. On the high density DNA Probe arrays of Affymetrix synthetic DNA fragments are sythesized on the GeneChip© (Lipshutz et al., 1999). Here, every gene is represented by one or more probe sets that in turn consists of up to 20 oligonucleotide probe pairs. Each probe pair is divided in two probe cells, and each probe cell consists of approximately 10^7 identical 25-mer oligonucleotides. In the first probe cell, the so called Perfect Match (PM), the nucleotide stretches matches perfectly with the one of the gene. The second probe cell is a kind of control of hybridisation signal where the sequence contains a non-matching nucleotide (MM). The 11–20 probe pairs of a probe set are randomly distributed on the array to avoid spatial effects. The values for each probe on an array is provided in the so called Cel-File and needed to be combined in a suitable way to one value for each probe set.
In addition, the *low-level analysis* ensures the comparability of values from different samples. It proceeds in four steps: Background correction, normalisation, probe specific background correction and finally combining to one value per chip and probe set. Many methods for this steps have been proposed in the literature (e.g. Lazaridis et al., 2002; Bolstad et al., 2003). We use the RMA (robust multi-array average) method (Irizarry et al., 2003) as implemented in the `R` (R Core Team, 2013) package `"affy"` (Gautier et al., 2004) available on Bioconductor (*http://www.bioconductor.org/*).

2.2.1 Breast cancer datasets

In this thesis, three breast cancer datasets where gene expression of node-negative (N0) patients is measured on Affymetrix GeneChip© HG U133A are considered. For all three cohorts, information on metastasis-free survival times are available.

The first data set derives from a population-based cohort study consisting of 200 patients consecutively treated at the Department of Obstetrics and Gynecology of the Johannes Gutenberg University Mainz between 1988 and 1998 (Schmidt et al., 2008). Therefore, this data set is referred to as "Mainz cohort". After surgical intervention in form of a modified radical mastectomy (75 patients) or a breast-conserving surgery followed by irradiation (125 patients), none of the women received a systemic therapy. During surgery no patient showed evidence of regional lymph node nor distant metastasis. From the original pathological report established prognostic factors like histological grade, tumor size, and steroid receptor status as well as data of the age at diagnosis date were recorded. The median age of the 200 patients at surgery was 60 years (34 to 89 years). The median follow-up time was 7 years and 8 months. Data is acessible through Gene Expression Omnibus (GEO, Edgar et al., 2002), accession number GSE11121.

Frozen tissue samples were selected from the tumor bank of the Erasmus Mediacal Center in Rotterdam (Netherlands). All 286 patients were diagnosd with node-negative breast cancer and surgically treated between 1980 and 1995 with a breast-conserving therapy or modified radical mastectomy, 219 and 67 patients, respectively. As in the Mainz cohort, none of them received any sytemic neoadjuvant or adjuvant chemotherapy (Wang et al., 2005), but 248 patients received a radiotherapy. Median patient age at surgery was below the median age in the Mainz cohort with 52 years (range from 26 to 83 years). Median follow-up time was about 7 years and 2 months, examinations were dated every 3 months in the first two years after surgery, every 6 months up to the fifth year and afterwards annual examinations were scheduled. This data set can be found by accession number GSE2034 in the GEO data base.

The third data set consists of two cohorts reported through accession numbers GSE6532

and GSE7390 in the GEO data base. The TRANSBIG cohort formed a collection of untreated node-negative breast cancer samples from patients of five European centers: Institut Gustave Roussy in Villejuif, France; Karolinska Institute in Stockholm, Sweden; Center René Huguenin in Saint-Cloud, France; Guy's Hospital in London, United Kingdom, and John Radcliffe Hospital in Oxford, United Kingdom (Buyse et al., 2006). Further criteria of inclusion were that patients were diagnosed between 1980 and 1998 with node-negative breast cancer with tumor size $\leq$ 5 cm and without previous malignancies or bilateral synchronous breast carcinomas, had not received any systemic adjuvant therapy, and were younger than 61 years at diagnosis. The latter fact was not consequently complied, but it makes the TRANSBIG cohort the study with with the youngest median age of 49 years (range from 24 to 73 years). It has the longest median follow-up time with about 13 years and 7 month. Assessment of grading according to Elston and Ellis was missing for 15 patients, while the grading for all 200 patients of the Mainz cohort was available. The data set of the Rotterdam cohort contains no grading information at all.

2.2.2 Non-small cell lung cancer datasets

The second collection of datasets used in this thesis consists of patients diagnosed with non-small cell lung cancer (NSCLC). Again, we require information about event times and gene expression data measured on Affymetrix HG U133A with 22 283 or HG U133 Plus 2.0 array. With the latter, 54 675 probe sets are measured. Except for 6 features, all probe sets of the HG U133A can be found on the HG U133 Plus 2.0 array. Therefore, the overlap of 22 277 probe sets is considered for the analysis performed in this thesis. For all cohorts, overall survival or censoring times are available.

The first NSCLC dataset, which is denoted as our basic cohort and is used for generating hypotheses, derives from patients operated in Uppsala University hospital in the years 1995–2005 with primary lung tumors and reported to the Uppsala-Örebro Regional Lung Cancer Registry (Botling et al., 2013). Further criteria of inclusion were that fresh frozen tumor tissue must be available in the Uppsala frozen tissue Biobank, the tumor must be larger than 5 mm, it must be confirmed as adenocarcinoma, squamous or large cell carcinoma/other NSCLC (NSCLC not otherwise specified (NOS)), and the fraction

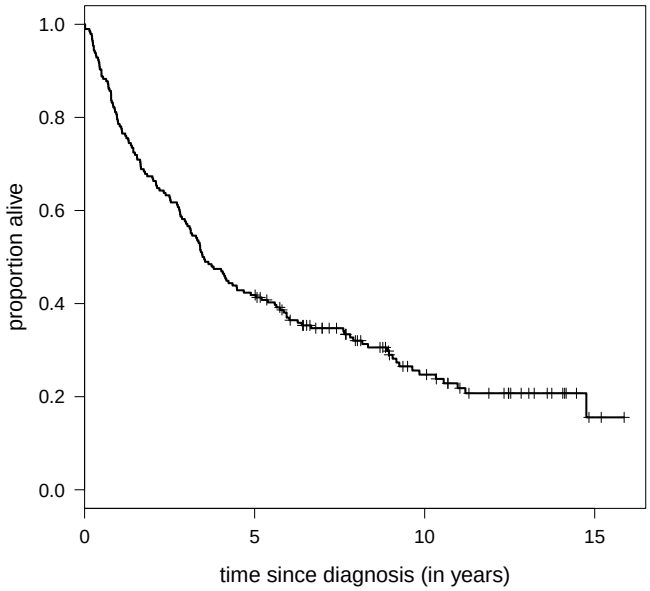

Figure 1: Kaplan-Meier curve for overall survival of the Uppsala NSCLC cohort.

of tumor cells must be above 50%. Patients who had received a neoadjuvant treatment were excluded. In total, 196 patients were included in the study. Information on several clinical and histopathological variables are available through the Uppsala-Örebro Regional Lung Cancer Registry like sex, age at diagnosis, performance status according to WHO (Oken et al., 1982) and the reports of the pathologists of the Uppsala University hospital on TNM staging. 106 (54.1%) patients were diagnosed with adenocarcinomas, and 66 and 24 with squamous and large cell carcinomas/NOS, respectively. This approximately complies with the proportions of all NSCLC cases (cf. Section 2.1). The median follow-up time was about 3 and a half years, no observation was censored within five years after surgery as we can see in Figure 1. Here, the Kaplan-Meier curve (e.g. Klein and Moeschberger, 2003) for all patients of the Uppsala lung cancer cohort is shown. Although patients with late stage NSCLC are usually not operated, 10 patients diagnosed with stage IIIb or even stage IV were operated and included in this study. This dataset, also referred to as the Uppsala cohort, is available via Gene Expression Omnibus, accession number GSE37745.

dataset	AC	SCC	other histology	total	Affymetrix array
GSE37745	106	66	24	196	HG U133 Plus 2.0
Jacob	448	-	-	448	HG U133A
GSE4573	-	130	-	130	HG U133A
GSE31547	30	-	-	30	HG U133A
GSE3141	58	52	-	110	HG U133 Plus 2.0
GSE29013	30	25	-	55	HG U133 Plus 2.0
GSE31210	204	-	-	204	HG U133 Plus 2.0
GSE19188	40	24	18	82	HG U133 Plus 2.0
GSE14814	28	52	10	90	HG U133A

Table 2: Overview of considered non-small cell lung cancer datasets.

In addition to the Uppsala NSCLC cohort where gene expression is measured on HG U133 Plus 2.0 arrays, 8 more datasets are considered for validation in this thesis. An overview of these cohorts is given in Table 2.

The dataset "Jacob" provided by Shedden et al. (2008) on the caArray platform (*https://array.nci.nih.gov/caarray/*) of the National Cancer Institute with experiment identifier "jacob-00182" consists of 448 early-stage (Ib and II) lung adenocarcinomas. Extended information on clinical data like age at diagnosis, sex, but also additional event times such as progression-free survival is available for these samples collected from the treatment institutions University of Michigan Cancer Center, Moffitt Cancer Center, Memorial Sloan-Kettering Cancer Center, and the Dana-Farber Cancer Institute, USA. Gene expression of this multi-center cohort was measured on Affymetrix HG U133A arrays.

From the Gene Expression Omnibus platform the dataset with accession number GSE31547 was used in addition. Gene expression was also measured on the Affymetrix HG U133A chip and the cohort consists also of the histological subtype of adenocarcinomas. This dataset is contributed by Girard from the Hamon Center for Therapeutic Oncology Research at Southwestern Medical Center, Dallas, USA, but the study had not been published yet. The original dataset as stored in GEO with accession number GSE31547 contains 30 primary lung adenocarcinomas and 20 adjacent normal lung controls. In this thesis, only the 30 adenocarcinomas are used for which additional information on clinical parameters are available. One more dataset that consists only of adenocarcinomas is considered in this thesis. GSE31210 (Okayama et al., 2012) contains required information on overall survival times of patients as well as gene expression mea-

surements on HG U133 Plus 2.0 arrays of 204 patients that were diagnosed with stage I or II between 1998 and 2008 at the National Cancer Center Hospital, Japan. These patients did not receive any neoadjuvant therapy nor a postoperative chemotherapy and/or radiotherapy after complete resection of tumor tissue.
In contrast, the study with accession number GSE4573 consists of 130 samples from 129 patients diagnosed with squamous cell carcinomas only (Raponi et al., 2006). For one patient two samples from different areas of the same tumor were taken and microarrays were prepared. Common clinical information of the patients diagnosed with stage Ia to IIIb and received surgically resection of tumor are available.
Besides these datasets containing only one histological subtype, we consider four more mixed cohorts. GSE3141 (Bild et al., 2006) has nearly balanced numbers of patients with adenocarcinomas and squamous cell carcinomas as well as GSE29013 (Xie et al., 2011). In contrast to GSE3141, the latter contains many clinical parameters such as TNM stage and age at diagnosis, but the microarrays are made of formalin-fixed paraffin-embedded (FFPE) samples which has disadvantages in terms of RNA degradation (Zhu et al., 2010). GSE19188 (Hou et al., 2010) contains tumor samples of 40 adenocarcinomas, 24 squamous cell, and 18 large cell carcinomas. Also GSE14814 (Zhu et al., 2010) provides data of 10 large cell carcinomas in addition to the two most frequent histological subtypes of early-stage patients. One criterion of inclusion here was that tumor cellularity was higher than 20% which is considerably low compared for example with the tumor cell fraction of at least 50% in the Uppsala NSCLC cohort.

3 Validation approaches for high-dimensional genetic data

Validation has become a major issue in biological analysis especially since high-dimensional data are available. Due to the enormous numbers of genes that are analyzed, the chance of observing significant results just by chance is high. Hence, adjustment of the α-level or rather p-values is required. Thereby the global α-level is controlled.
In this thesis we present different approaches for the validation of significant features on high-dimensional datasets applied to the cohorts introduced in Section 2.2. Significant features can be obtained e.g. from two sample t-tests that compare the means of gene expression values between two groups of patients or from Wald tests that identify genes correlated with survival. The performance of the methods will be analyzed by a simulation study to assess the number of false positive and false negative features.

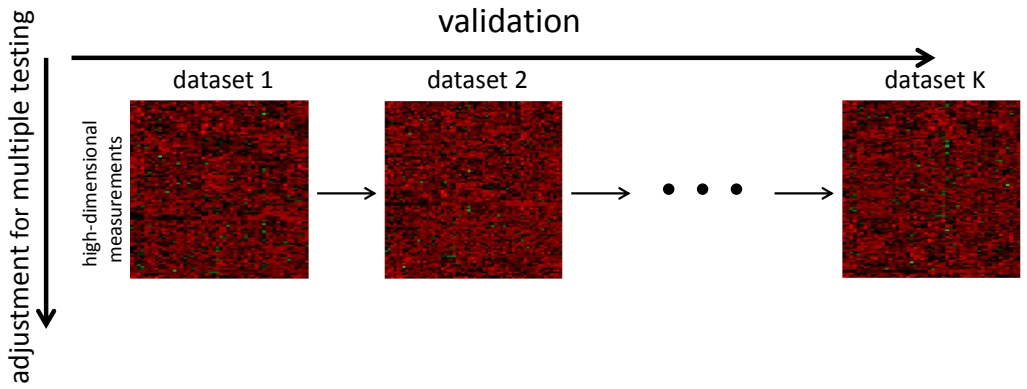

Figure 2: Illustration of validation and simultaneous adjustment for multiple testing on $K > 1$ high-dimensional datasets.

Actually, for validation on other datasets, adjustment must be performed on every considered cohort. Whether the true significant features stand out in another dataset depends on the quality of the studies. We assume quality to be a composition of sample size of a study and the underlying noise. The noise might originate from differences in specific technical procedures performed in a specific medical center, to differences between compositions of the samples in different medical centers, or maybe unconsidered biological differences of the study individuals.
Though, to identify the true positive, which means true significant, genes will be quite difficult after strict adjustment due to differences between the studies. Hence less stringent adjustment methods are required. The issue of validation and simultaneous need of adjustment for multiple testing is illustrated in Figure 2. Due to the large number of tested genes the number of erroneously rejected hypotheses must be controlled and concurrently we aim for a validation of findings on other datasets.

The combination of two or more datasets to one combined cohort and subsequent analyzing and trying to confirm the results e.g. by cross-validation seems to be a poor concept. In most cases a batch effect occurs. An example is given in Figure 3. Here, we see the expression values examplified for the first common probe set "1007_s_at" representing gene "DDR1" in the list of the nine non-small cell lung cancer datasets introduced in Section 2.2.2. for every patient. The expression values of each cohort are painted in an distinct color. We recognize that values of one cohort are often strictly separated from the values of the others. Combining those values without batch normalization will bias the result of most analyses.
A better approach for combined results is the ordinary meta-analysis. One advantage which might be concomitant a disadvantage is that many small signals may be sufficient for a significant result in a meta-analysis. Besides it takes all studies into account simultaneously. However, our focus is on another strategy: We have a basic dataset that we use to identify interesting genes. Afterwards these genes should be validated on other datasets. Though, we have to find a tradeoff between validation and strict adjustment for multiple testing to receive as many relevant genes as possible and simultaneously eliminating all false positive, i.e. non-relevant genes.
In the following section we give a short introduction to multiple testing and particularly the adjustment of p-values by controlling the False Discovery Rate (FDR). Further some

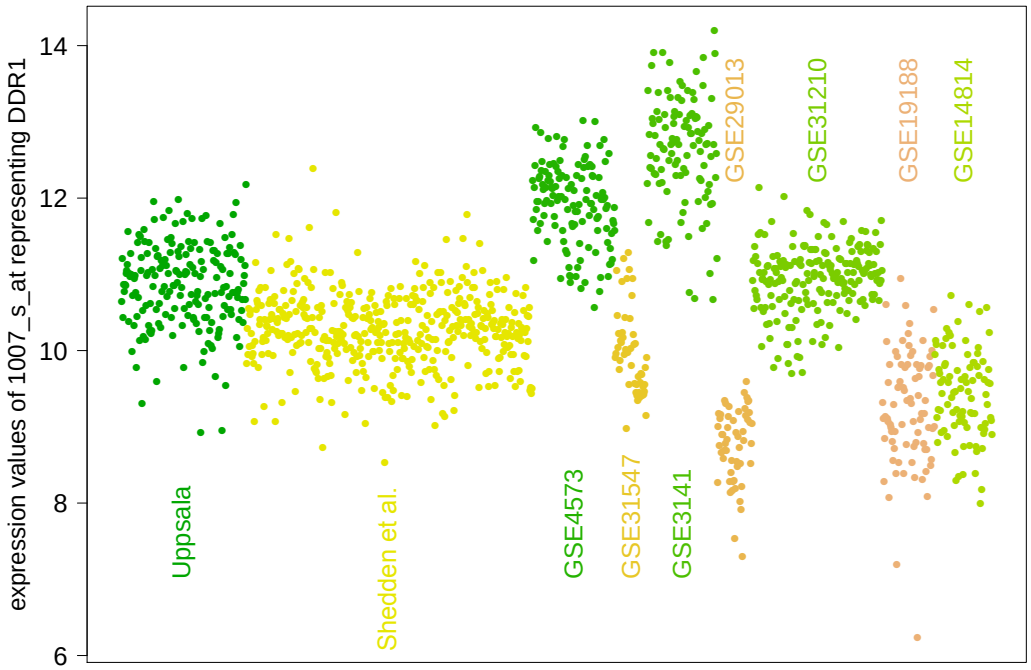

Figure 3: Expression values of probe set "1007_s_at" representing gene "DDR1" of all nine non-small cell lung cancer cohorts.

methods for survival analysis and meta-analysis used in this chapter are described. A sequential validation strategy is proposed in Section 3.2. We demonstrate the use of this procedure concerning the elimination of false positive results and simultaneously maximizing the power of the procedure in a simulation study, as well as the use in assessing the quality of datasets. We also apply the strategy to several cancer datasets that are introduced in Section 2.2. Another validation approach is introduced in Section 3.3, and performed on non-small cell lung cancer datasets. The special feature of this strategy is the 2-step procedure of first on unadjusted screening and than a meta-analysis for validation. In addition, an ordinary meta-analysis is performed on the lung cancer datasets. In Section 3.5, the results of the three approaches are compared and discussed.

3.1 Methods

3.1.1 Multiple testing

If a statistical test is performed, two outcomes are possible: The null hypothesis (H_0) can be rejected, i.e. the test is significant, or there is no sufficient proof against H_0 and the test is not declared to be significant. On the other hand the null hypothesis can be true or false. If H_0 is true and it is not rejected we obtain a correct decision, just as if the null hypothesis is not true and the test is significant. Hence, two (types of) errors can be committed (Dudoit and van der Laan, 2008). We call a decision *type I error* or *true positive* if a true null hypothesis is rejected and a *type II error* or *false negative* is committed if a non-true null hypothesis is not rejected. In modern testing theory the main focus is on the type I error that is controlled by a level α, i.e. $P(\text{rejecting } H_0 | H_0 \text{ is true}) \leq \alpha$ and we merely try to minimize the type II error given a fixed value for α.

Number of	**Number not rejected**	**Number rejected**	**Total**
true H_0	A	V	g_0
non-true H_0	F	B	g_1
	$g - R$	R	g

Table 3: Number of correct decisions and type I and type II errors committed in multiple testing of g hypotheses (reproduced from Benjamini and Hochberg, 1995).

In many fields like the analysis of gene expression data thousands of hypotheses are tested simultaneously. If we test lots of hypotheses at a specified significance level the probability of committing type I errors increases with the number of hypotheses. Therefore, an adjustment for multiple testing is required to control the type I error rate. Let us assume we want to test g null hypotheses H_0^i, $i = 1, \ldots, g$ of which g_0 are true and $g_1 = g - g_0$ hypotheses are false. Both parameters, g_0 and g_1 are unknown. The number of rejected and not rejected null hypotheses are treated as random variables denoted by R and $g - R$, respectively. Table 3 (cf. Benjamini and Hochberg, 1995) summarizes the numbers of true/non-true and (not) rejected hypotheses. A, F, V, and B are unobservable random variables where V are the false positives or type I errors and F denotes

the number of false negatives or rather type II errors. As mentioned above, it is the philosophy of significance testing to control the number of type I errors. In the following the two most common type I error rates are introduced as well as one multiple testing procedure for each error rate.

The family-wise error rate

We define the family-wise error rate (FWER) as the probability of rejecting at least one null hypothesis when it is true, i.e. committing at least one type I error under all decisions (Dudoit et al., 2003):

$$\text{FWER} = P(V \geq 1).$$

To control the FWER on a global level α the p-values have to be adjusted. One common procedure that is used in this thesis is the *Holm procedure* (Holm, 1979) which is a further development of the Bonferroni procedure (Dudoit and van der Laan, 2008). Compared to Bonferroni as a single-step procedure the Holm procedure adjusts the p-values step-down which is an advantage because it is less conservative.
Let $p_1 \leq p_2 \leq \ldots \leq p_g$ denote observed and ordered raw p-values and let $H_0^1, H_0^2, \ldots, H_0^g$ be the corresponding null hypotheses. We define

$$i^\star = \min \left\{ i : p_i > \alpha / \left(g - i + 1 \right) \right\}.$$

If such $i^\star$ exists, reject all null hypotheses H_0^i, with $i = 1, 2, \ldots, (i^\star - 1)$. When no such $i^\star$ exists, all g null hypotheses are rejected. The adjusted p-values can be calculated by

$$\tilde{p}_i = \max_{l=1,\ldots,i} \left\{ min \left(\left(g - l + 1 \right) p_l, 1 \right) \right\}.$$

Afterwards, these adjusted p-values can be compared with the local level α.

The false discovery rate

The false discovery rate (FDR) described by Benjamini and Hochberg (1995) is defined as the expected proportion of type I errors among the reject hypotheses:

$$\text{FDR} = E(O)$$

where O is defined by $O = \frac{V}{R}$. This error rate tolerates some type I errors, but the proportion of errors among all rejected hypotheses is controlled. If all null hypotheses are true, the false discovery rate is equilavent to the family-wise error rate because $b = 0$, where b denotes an observation of B and $V = R$. Let v denote an observation of V and define $\frac{0}{0} := 0$. If the numbers of true as well as false null hypotheses are greater then 0, we have to consider two cases:

1. $v = 0$ (no H_0 is rejected) $\Rightarrow O = \frac{V}{R} = 0 \Rightarrow P(V \geq 1) = E(O)$

2. $v > 0 \Leftrightarrow v \geq 1$ (at least one H_0 is misleadingly rejected):
   $\frac{v}{r} \leq 1 \Rightarrow I_{(V \geq 1)} \geq \frac{V}{R} = O \Rightarrow P(V \geq 1) \geq E(O)$

Thus, by controlling the FWER, the FDR is also controlled in the weak sense.
Benjamini and Hochberg (1995) provide a procedure for controlling (only) the FDR for independent tests which is less stringent and power is gained, i.e. we yield a smaller number of type II errors. Other procedures for the control of the FDR under certain dependence structures are provided for example by Benjamini and Yekutieli (2001). Therefore, let $p_1 \leq p_2 \leq \ldots \leq p_g$ denote observed and ordered raw p-values as in the last paragraph and we define

$$i^{\star\star} = \min \{i : p_i \leq (i/g)\, \alpha\}\,.$$

If all hypotheses H_0^i, with $i = 1, 2, \ldots, i^{\star\star}$, were rejected the FDR is controlled at level α. If no such $i^{\star\star}$ exists, we reject no hypothesis. The corresponding adjusted p-values can be calculated as follows:

$$\tilde{p}_i = \min_{l=i,\ldots,g} \left\{ \min \left(\frac{g}{l} p_l, 1 \right) \right\}.$$

These adjusted p-values can be compared with the local level α as for the Holm procedure.

3.1.2 Cox proportional hazards model

Survival analysis deals with the problem of analyzing time-to-event data. In our case of biological data, events could be, for example, death, recurrence of cancer, or distant metastasis. Not for all patients the event of interest can be observed in the period fixed. If the event has not occured until at the end of the period or a patient drops out of the study for other reasons, the corresponding observation will be *(right) censored*. High censoring rates are often seen in cancer data. But instead of eliminating all patients with missing endpoints, the censored data is integrated in survival analysis. Even if we do not know the exact event time of a censored patient, we know that he or she has not seen the event until his or her drop-out. An important assumption for the following methods of survival analysis is that censoring is independent from the event of interest. Before we specify the Cox proportional hazard model, the notation and some basic quantities will be introduced.
Let T be the time from a starting point $t = 0$ to the event of interest, i.e. a non-negative random variable, and $f(t)$ its density with distribution function $F(t)$ (see Klein and Moeschberger, 2003). The *survival function* is the probability of observing the event after time point t, defined as

$$S(t) := P(T > t).$$

Since T is a continuous random variable $S(t)$ is a continuous, strictly decreasing function with $\lim_{t\to 0} S(t) = 1$ and $\lim_{t\to\infty} S(t) = 0$ and moreover the complement of the distribution function $S(t) = 1 - F(t) = 1 - P(T \leq t)$. Alternatively, the survival function can be specified by the integral of the density, $f(t)$, i.e.

$$S(t) = P(T > t) = \int_t^\infty f(u)\, du.$$

The probability that an individual experiences the event in the next instant, conditional on that the person was event-free until t, is called risk function or *hazard rate (function)*

$h(t)$:

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t},$$

where $h(t) \geq 0, \forall\, t \in [0, \infty]$. The hazard rate is connected to the survival function as follows:

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T > t)} = \frac{f(t)}{S(t)}.$$

Beside the plain hazard rate, the probability of an individual experiences an event in the next instant time, conditional on survival time t with a specific value of a covariate is considered. Possible candidates for covariates that might have impact on the probability that an individual experiences an event in the next instant may be e.g. gender, a special treatment, or gene expression. This conditional probability can be described by the *Cox proportional hazards model* (Cox, 1972).

Let the data consist of n samples. For each sample or patient the triple $(T_j, \delta_j, \mathbf{Z}_j)$, $j = 1, \ldots, n$ is available, where T_j is the time that patient j spent in the study, δ_j the indicator whether the $j-$th patient experienced the event ($\delta_j = 1$) or was censored ($\delta_j = 0$), and $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jp})^{\text{`}}$ is a vector of p covariates for individual j. Cox (1972) defines the basic model as follows:

$$h\left(t|\mathbf{Z}\right) = h_0(t) \cdot \exp\left(\boldsymbol{\beta}^{\text{`}}\mathbf{Z}\right) = h_0(t) \cdot \exp\left(\sum_{k=1}^{p} \beta_k Z_k\right),$$

where $h_0(t)$ is an arbitrary baseline hazard rate (function), and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\text{`}}$ is a parameter vector. The Cox proportional hazards model is a semiparametric model because it consists of a nonparametric part, the baseline hazard rate, and a parametric part since the effect of the covariates is assumed to be parametric. An important assumption of the model is couched in its name: the proportionality of hazard rates. If we have two patients with covariate values $\mathbf{Z_1}$ and $\mathbf{Z_2}$, then the ratio of the two hazard rates is constant, i.e. independent of time t:

$$\frac{h(t|\mathbf{Z_1})}{h(t|\mathbf{Z_2})} = \frac{h_0(t) \cdot \exp\left(\boldsymbol{\beta}^{\text{`}}\mathbf{Z_1}\right)}{h_0(t) \cdot \exp\left(\boldsymbol{\beta}^{\text{`}}\mathbf{Z_2}\right)} = \frac{h_0(t) \cdot \exp\left(\sum_{k=1}^{p} \beta_k Z_{1k}\right)}{h_0(t) \cdot \exp\left(\sum_{k=1}^{p} \beta_k Z_{2k}\right)} = \underbrace{\exp\left[\sum_{k=1}^{p} \beta_k \left(Z_{1k} - Z_{2k}\right)\right]}_{\text{constant in } t}.$$

This ratio is called relative risk or *hazard ratio* (HR) and describes the relative risk

of a patient with covariate vector $\mathbf{Z_1}$ experiencing the event of interest in the next instant compared to a patient with covariate vector $\mathbf{Z_2}$. In this thesis, we restrict on the analysis of univariate Cox proportional hazards models. If so, and the covariate has two categories, e.g. treatment and control, $\exp(\beta_1) = \frac{h(t|Z_1)}{h(t|Z_2)}$ describes the risk for the occurence of the event of a patient who received the treatment relative to the risk for an individual of the control group. A similar interpretation is possible for hazard ratios of multivariate Cox proportional hazards models, however, the values of all other covariates that are considered in the model must be the same in both groups.
The parameter vector $\boldsymbol{\beta}$ can be estimated by a Maximum Likelihood approach. Suppose that there are no ties between the observed event times, and let $t_1 < t_2 < \ldots < t_D$ denote the ordered event times with corresponding k-th covariate $Z_{(i)k}, i = 1, \ldots, D$ for the patient with event time t_i. The risk set at time point t_i, denoted by $R(t_i)$, is defined as the set of all individuals who have not seen the event till time t_i or dropped out of the study, i.e. they are still under risk and might see the event in the future. The *partial likelihood* is based on hazard rates and is composed of the information of an individual patient i experiencing the event (in the numerator) and the information about all patients that are still under study for every time point t_i (in the denominator) as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left[\sum_{k=1}^{p} \beta_k Z_{(i)k}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k Z_{jk}\right]}.$$

The name "partial" likelihood refers to the fact that this expression ignores the actual event times but takes the order of the latter into account. Instead of maximizing the partial likelihood we maximize the log partial likelihood which can be written as

$$\begin{aligned} LL(\beta) &= \ln\left(\prod_{i=1}^{D} \frac{\exp\left[\sum_{k=1}^{p} \beta_k Z_{(i)k}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k Z_{jk}\right]}\right) = \sum_{i=1}^{D} \ln\left(\frac{\exp\left[\sum_{k=1}^{p} \beta_k Z_{(i)k}\right]}{\sum_{j \in R(t_i)} \exp\left[\sum_{k=1}^{p} \beta_k Z_{jk}\right]}\right) \\ &= \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k Z_{(i)k} - \sum_{i=1}^{D} \ln\left[\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)\right]. \end{aligned}$$

Solving the equation system $\frac{\partial}{\partial \beta_k} LL(\boldsymbol{\beta}) = 0, \forall k = 1, \ldots, p$ leads to the maximum likelihood estimation $\hat{\boldsymbol{\beta}}$ of the regression coefficients $\boldsymbol{\beta}$.
After the estimation of the parameter vector $\hat{\boldsymbol{\beta}}$, the baseline hazard rate $h_0(t)$ is esti-

mated by the Breslow estimator

$$\hat{h}_0(t) = \prod_{t \leq t_i} \left(\frac{d_i}{\sum_{j \in R(t_i)} \exp\left(\hat{\beta}' Z_j\right)} \right)$$

with d_i as the number of events at time point t_i. As mentioned before, we assume to have no ties between the event times since time is continuous and therefore we have $d_i = 1 \; \forall i = 1, \ldots, D$.

Finally, when the Cox proportional hazards model with the chosen variables is estimated, it will be of interest which covariates have a significant effect on the individual risk. In this thesis we restrict to univariate Cox proportional hazards models, as mentioned above. Therefore, testing the hypothesis $H_0 : \beta_1 = 0$, that is equivalent to $H_0 : \text{HR} = 1$, vs. $H_1 : \beta_1 \neq 0$ is sufficient to answer the question if the covariate $\mathbf{Z_1}$ has a significant influence on the event time, e.g. survival. The statistic of the *local (Wald like) test* (see Wald, 1943) that is used in this thesis is defined as

$$\chi^2_W = \frac{\hat{\beta}_1}{\sqrt{SE(\hat{\beta}_1)}},$$

where the standard error $SE(\hat{\beta}_1)$ is derived by i-th diagonal element of $\left[-\frac{\partial^2}{\partial\beta_j \partial\beta_k} LL(\boldsymbol{\beta})\right]_{j,k=1,\ldots,p}$ or rather $-\frac{\partial^2}{\partial^2 \beta_1} LL(\beta_1)$ in the special case of $p = 1$. The null-hypothesis is rejected if $\chi^2_W > \chi^2_{p,1-\alpha}$, where $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with $p = 1$ degrees of freedom.

3.1.3 Meta-analysis

Since meta-analysis was first defined by Glass (1976) as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings", it has become indispensable in todays clinical research. Especially in the context of evidence-based medicine meta-analyses are essential for the highest level of evidence (Evidence-Based Medicine Working Group, 1992). Many statistical methods for conducting meta-analyses were developed and refined (see Sutton et al., 2000, or Stangl

and Berry, 2000). We restrict to one or rather two mainstream statistical approaches (cf. Whitehead, 2002).
A main focus of controversy is the choice between the *fixed effect* and the *random effects model* that enable an estimate of the overall effect of interest. In the fixed effect model the effect of every study is considered to be out of the same (normal) distribution, while in the random effects model an additional between study effect is assumed. The biggest disadvantages of the fixed effect model are that it does not hold under *heterogeneity* (see e.g. Schumacher and Schulgen, 2006) and it holds only for the particular studies included in the meta-analysis, i.e. a generalization of the results is quite problematic. The random effects model consideres this issue and is more generazible. However, it must be taken into account that often the trials included in the meta-analyses are not representative for the total polulation. Moreover, if only a small number of studies are taken into consideration, the between-study effect fitted by a random effects model might be unreliable. Before both models are briefly introduced, we define some notation.
Let K be the number of independent studies and θ the true effect of interest that shall be estimated by $\hat{\theta}$. θ_k, $k = 1, \ldots, K$ denotes the single study effect of the k-th study that is estimated by $\hat{\theta}_k$.

The fixed effect model

In Whitehead (2002) the general fixed effect model is defined by

$$\hat{\theta}_k = \theta + e_k, \qquad \theta_k = \theta \quad \forall k = 1, \ldots, K$$

for $k = 1, \ldots, K$, where the errors e_k are realizations of normally distributed random variables $\epsilon_k \sim N(0, \xi_k^2)$. It follows that

$$\hat{\theta}_k \sim N(\theta, \xi_k^2).$$

We treat the variance of $\hat{\theta}_k$, $\text{var}\left(\hat{\theta}_k\right)$ as if it were the true variance (ξ_k^2) and denote $w_k = \frac{1}{\text{var}\left(\hat{\theta}_k\right)}$ as the inverse estimated variance of $\hat{\theta}_k$. Then we assume that

$$\hat{\theta}_k \sim N(\theta, w_k^{-1}).$$

The overall (fixed) effect θ is estimated by a weighted sum of the single study effects:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \hat{\theta}_k w_k}{\sum_{k=1}^{K} w_k}.$$

A calculation of the standard error of $\hat{\theta}$ can be obtained as follows

$$SE(\hat{\theta}) = \sqrt{\frac{1}{\sum_{k=1}^{K} w_k}}.$$

Thus, an appromimate 95% confidence interval (CI) for the overall effect is given by

$$\left[\hat{\theta} - 1.96\sqrt{\frac{1}{\sum_{k=1}^{K} w_k}}\,;\, \hat{\theta} + 1.96\sqrt{\frac{1}{\sum_{k=1}^{K} w_k}}\right].$$

Using a *test for heterogeneity* we are able to test the hypothesis $H_0 : \theta_1 = \theta_2 = \ldots = \theta_K$ against $H_1 : \theta_k \neq \theta_l$ for at least one pair (k, l), $1 \leq k, l \leq K$, $k \neq l$ (see Whitehead, 2002). The test statistic is given by

$$Q = \sum_{k=1}^{K} w_k \left(\hat{\theta}_k - \hat{\theta}\right)^2.$$

If the null-hypothesis is true, i.e.
all study effects are homogeneous, Q follows a χ distribution with $(K-1)$ degrees of freedom. Thus, we can reject the null-hypothesis if $Q > \chi^2_{K-1,1-\alpha}$, where $\chi^2_{K-1,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with $(K-1)$ degrees of freedom.

The random effects model

In a random effects model we assume the single study effects $\theta_1, \theta_2, \ldots, \theta_K$ to follow a normal distribution with mean θ and variance τ^2. The general random effects model

is defined as follows:

$$\hat{\theta}_k = \theta + \nu_k + \epsilon_k,$$

for $k = 1, \ldots, K$. Here, ν_k are random effects (or between study effects) with $\nu_k \sim N(0, \tau^2)$ and ϵ_k is a normally distributed error with $\epsilon_k \sim N(0, \xi_k^2)$ as before. Assuming that ν_k and ϵ_k are independently distributed, it follows that

$$\hat{\theta}_k \sim N\left(\theta, \xi_k^2 + \tau^2\right).$$

The between study variance τ^2 can be estimated by $\hat{\tau}^2$ from the data using the method of moments described by DerSimonian and Laird (1986):

$$\hat{\tau}^2 = \frac{Q - (K-1)}{\sum_{k=1}^K w_k - \frac{\sum_{k=1}^K w_k^2}{\sum_{k=1}^K w_k}}.$$

With this estimator we made the assumption that

$$\hat{\theta}_k \sim N\left(\theta, w_k^{-1} + \hat{\tau}^2\right).$$

We define $w_k^\star = \left(w_k^{-1} + \hat{\tau}^2\right)^{-1}$ and it follows that

$$\hat{\theta}_k \sim N\left(\theta, (w_k^\star)^{-1}\right).$$

As in the fixed effect model, we treat $(w_k^\star)^{-1}$ as if it were the true variance of $\hat{\theta}_k$ and we obtain

$$\hat{\theta^\star} = \frac{\sum_{k=1}^K \hat{\theta}_k w_k^\star}{\sum_{k=1}^K w_k^\star}$$

as maximum likelihood estimate of θ. Analogously to the fixed effect model, the standard error is given by

$$SE(\hat{\theta}^\star) = \sqrt{\frac{1}{\sum_{k=1}^K w_k^\star}},$$

which leads to an approximate confidence interval for the effect of interest θ:

$$\left[\hat{\theta}^\star - 1.96\sqrt{\frac{1}{\sum_{k=1}^K w_k^\star}}\,;\, \hat{\theta}^\star + 1.96\sqrt{\frac{1}{\sum_{k=1}^K w_k^\star}}\right].$$

The weights w_k and $w_k^\star$ of the fixed effect and the random effects model, respectively, will not differ much if the estimate of the between study effect $\hat{\tau}^2$ is close to zero. The obtained estimates of the overall effect, the standard errors as well as the confidence intervals will be hardly the same in this case. However, if $\hat{\tau}^2$ is large the standard error will be larger for the random effects model and with it the confidence interval for θ. Besides, the estimate of the latter model will move closer to the arithmetic mean. The amount depends on the study with the largest weights in the fixed effect model.

3.2 Sequential validation strategy

Clearly discriminated from the concept of the ordinary meta-analysis we propose a stepwise validation procedure (Lohr et al., 2012). We assume that some characteristic shall be tested genome-wide, e.g. differential expression among two conditions or patient groups by conducting t-tests or the correlation with a specific event time by a Wald tests, on serveral datasets. Here, we restrict on gene expression data, but the algorithm is also applicable for other types of high-dimensional data, e.g. SNP (single-nucleotide polymorphisms) data. Suppose K datasets are available and g genes are measured in each cohort. In total we obtain $K \cdot g$ p-values that require a reasonable adjustment.
We propose the following algorithm that is briefly explained for $K = 3$ datasets. Assume the datasets are ordered according to an arbitraty preference. For all genes in each datasets raw p-values have been calculated. In a first step, we adjust the p-values of the first dataset for multiple testing with method M_1. All genes related to p-values above the significance level α are excluded from further analysis. In a second step, the p-values belonging to the remaining genes are adjusted on the second dataset using method M_2. Usually the number of significant genes after adjustment for multiple testing in the first step is much smaller then the entire number of screened genes. For that reason the size of the adjusted p-values will increase. Again, we take the genes whose adjusted p-values are smaller then the α-level of the second dataset and reduce the third dataset to these genes. The remaining p-values of the third dataset are adjusted for multiple testing with method M_3 once more. Conducting this procedure the number of potential significant genes decreases from step to step. The general algorithm for $K > 3$ can be found in Figure 4 (see Lohr et al. 2012, p. 449).

Tuning at several points of the algorithm is possible. The significance level α can be selected as well as the methods of adjustment. We apply the algorithm to simulated data and to real cancer data. In simulation studies we have tested lots of combinations of these parameters and the results are summarized in Section 3.2.2. It will turn out that the best setting in simulations was to use the Benjamini-Hochberg procedure that controls the False Discovery Rate (FDR) as adjustment for multiple testing in combination with an significance level of $\alpha = 5\%$ in every step ($M1 = M2 = \ldots = M_K$). This setting is used to analyze three breast cancer datasets via the sequential validation algorithm.

Input: Raw p-values for K ordered datasets and g genes,
sorted sequence of length K of methods $M_1, \ldots, M_K$ for multiple testing,
critical level α for selection of significant features

Output: List of significant features genes after sequential validation in all K datasets

Algorithm:

Step 1. Adjust the raw p-values of the first dataset for multiple testing with method M_1 and select the remaining significant genes with p-value below α

step $k = 2, \ldots, K$.
Reduce dataset k to selected features, adjust the corresponding raw p-values for multiple testing with method M_k, and select the remaining significant features with p-value below α

step $K + 1$. Select final list of genes whose adjusted p-values are below α in step K.

Figure 4: Algorithm for stepwise p-value adjustment for K datasets (reproduced from Lohr et al., 2012).

3.2.1 Stepwise validation on breast cancer datasets

All three breast cancer cohorts of Mainz, Rotterdam, and TRANSBIG described in Section 2.2.1 were divided into two groups. A major challenge is the prediction of clinical outcome. Therefore we are looking for differentially expressed genes or rather probe sets in patients with and without metastastes as an indicator of recurrence free survial. The first group consists of patients that developed metastases, the second one of patients which had been observed for at least five years and did not develope any metastases. Patients that dropped out of the studies within five years without metastases were excluded from this analysis. Through this classification the metastases groups include 47, 107, and 72 patients and the metastases-free groups contain 136, 168, and 189 patients for the Mainz, Rotterdam, and TRANSBIG cohorts, respectively.

We apply t-tests to every probe set on all three datasets and apply the algorithm for every possible order of the studies. A basic dataset, to explore significant features in, is

sequence	1st step	2nd step	3rd step
Mainz → Rotterdam → TRANSBIG	32	8	2
Mainz → TRANSBIG → Rotterdam	32	12	2
Rotterdam → Mainz → TRANSBIG	133	43	24
Rotterdam → TRANSBIG → Mainz	133	45	24
TRANSBIG → Mainz → Rotterdam	0	0	0
TRANSBIG → Rotterdam → Mainz	0	0	0

Table 4: Numbers of significant probe sets in each step for the six validation sequences (using FDR for significance adjustment in every step and a α-level of 5%) for the three breast cancer cohorts (reproduced from Lohr et al., 2012).

not determined here, the purpose is to observe the results for all sequences. The numbers of significant probe sets in each step are shown in Table 4. We notice the number of significant probe sets after three steps only depends on the dataset that is adjusted first. Differences between the numbers after the second step are compensated in the third adjustment step. If we adjust the Rotterdam cohort first most probe sets are detected, while starting the adjustment with the dataset of TRANSBIG leads to no significant probe sets even in the first step. It is obvious that there must be crucial differences between the datasets regarding the size of signals. The TRANSBIG cohort seems to be associated with the smallest number of signals since the number of significant genes is regarded as an indication for the strength of signals. Only the less strong adjustment of the TRANSBIG cohort in a validation step, i.e. second or third step, yields significant genes.

The disagreement in the number of significant genes depending on the order of adjustment may be caused by two reasons: 1. the different sample sizes of the cohorts and/or 2. the underlyding noise in the data. While the TRANSBIG cohort is associated with the smallest number of significant probe sets and has a larger sample size than the Mainz dataset we expect the TRANSBIG cohort to be the dataset with the highest noise level. To draw any conclusions about the effect of sample sizes and noise levels, in the following the influence of varying noise levels and afterwards sample sizes on simulated data section is analyzed.

3.2.2 Performance check via simulation studies

Since real data is not suitable to analyze the properties of an algorithm, we performed extensive simulation studies. A selection of the results is presented in the following subsection. First we describe how the data is designed. As mentioned above we expect different underlying noise levels and sample sizes to be reasons for differing data qualities that are responsible for the varying numbers of significant probe sets. Therefore, we analyze the effect of varying noise levels first by simulating data with equal sample sizes. Afterwards we examine the additional influence of different sample sizes based on the breast cancer datasets.

Design of data

Data for two groups of patients, here for simplification we call them cases and controls, has to be generated. We assume the expression values of both groups to be normally distributed. For a subset of genes, that is assumed to be significant, the expression values of one group are shifted by realizations of an also normally distributed random variable C. Let n_1 and n_0 be the number of patients in the cases and the control group, respectively. We define g as the total number of measured genes and m, $m \leq g$ the number of differentially expressed genes between cases and controls. Let the baseline expression values for both groups $a_{i,j}$, $i = 1, \ldots, g$, and $j = 1, \ldots, (n_0 + n_1)$ be realizations of a normally distributed random variable A, where $A \sim \mathcal{N}(0, \sigma_k^2)$, $\sigma_k^2 > 0$ and let $c_{i,j}, i = 1, \ldots, m$, and $j = 1, \ldots, n_1$ be realizations of a normally distributed shift-variable C, $C \sim \mathcal{N}(\mu_s, \sigma_s^2)$, $\mu_s \in \mathbb{R}$, and $\sigma_s^2 > 0$. The matrix of the simulated expression values $\mathbf{x}$ is then composed of

$$x_{i,j} = \begin{cases} a_{i,j} + c_{i,j}, & i = 1, \ldots m,\ j = 1, \ldots, n_1, \\ a_{i,j}, & i = m+1, \ldots, g,\ j = n_0, \ldots, n_0 + n_1. \end{cases}$$

For the simulated data we determine some basic settings. The number of measured genes is set to $g = 20\,000$ and the number of differentially expressed genes is fixed to $m = 100$. In the following simulations we generate $c_{i,j}$ from a normal distribution with mean and variance 1, thus the effect in every study is assumed to be the same. We generate $K = 3$ datasets for each simulation and carry out the algorithm as described above. For the

three datasets we assume different quality levels which depend on the sample sizes and the underlying noise, that is variance. In the following section we assume equal sample sizes in the cases and the control group and all three datasets to be able to recognize the effect of different noise intensities σ_k^2 of the datasets. We denote the dataset with the lower quality, i.e. higher variance, by σ_{lq}^2, the one with medium quality by σ_{mq}^2, and the dataset with highest quality, i.e. lowest variance, by σ_{hq}^2.

Stepwise validation for simulated data with equal sample sizes

First, we set the number of patients per group to $n_1 = n_2 = 50$ in each dataset. Though, the qualities of the datasets are defined by their underlying noise level. Table 5 shows the results of the validation algorithm assuming a standard deviation of $\sigma_{hq} = 0.8$ on the high quality dataset (denoted by hq), a moderate noise level of $\sigma_{mq} = 1.1$ on the dataset with medium quality (mq), and a third low quality dataset (lq) with $\sigma_{lq} = 1.5$. In extended simulations these noise levels turned out to be realistic and the effect of the adjustment can be well studied. The median of the true positives as well as the median of the false negatives of 1000 simulations for all three steps are presented, on the left for adjustment with Bonferroni-Holm (Holm) and on the right side controlling the FDR with the Benjamini-Hochberg procedure at an α-level of 5%.

	$\sigma_{hq} = 0.8, \sigma_{mq} = 1.1, \sigma_{lq} = 1.5$					
	Holm			FDR		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $hq \to mq \to lq$	45 / 0	44.5 / 0	31 / 0	85 / 4	84 / 0	77 / 0
2: $hq \to lq \to mq$	45 / 0	31 / 0	31 / 0	85 / 4	78 / 0	77 / 0
3: $mq \to hq \to lq$	17 / 0	17 / 0	15 / 0	53 / 2	53 / 0	49 / 0
4: $mq \to lq \to hq$	17 / 0	15 / 0	15 / 0	53 / 2	49 / 0	49 / 0
5: $lq \to hq \to mq$	4 / 0	4 / 0	4 / 0	15 / 1	15 / 0	15 / 0
6: $lq \to mq \to hq$	4 / 0	4 / 0	4 / 0	15 / 1	15 / 0	15 / 0

Table 5: Median true positives/false positives in each step of the six adjustment sequences for simulated gene expression data (reproduced from Lohr et al., 2012).

Independent of the method of adjustment for multiple testing the median number of

true positives decreases from step to step as well as the median number of false positives. The optimal result would be to identify all true positives which means significant genes and reduce the number of false positives to zero, simultaneously. By performing the algorithm using Bonferroni-Holm correction no false positives are found even after the first adjustment step for all sequences of datasets in this setting. After the second adjustment step we observe no false positives even using the FDR. We get closest to our aim of 100 true positives and zero false positives if the dataset with the highest quality, i.e. here the lowest noise level, is adjusted first. Whether the dataset with medium or high underlying noise is adjusted next is less important because the difference in true positives after the second adjustment step is equalized after the third step.
We tested various parameter settings for the noise, adjustment methods (see Table 21 and 22 in the appendix), distributions for the shift-variable and (equal) number of observations as well as adapted settings on different kinds of data performing other tests. Beside simulating gene expression data and applying t-tests on it, we analogous generated SNP data with differential allele frequencies of 0.53 in the cases and 0.48 for the control group and apply χ^2-tests to test the hypothesis whether the allele frequencies are significantly different (see Lohr et al. 2012). Although, adapting the number of true positives after the first adjustment step by modulating the number of patients per group, we observe less true positives in the second step. We see, the type of data as well as the properties of the test procedure play a role for the amount of the true positive rate.
In summary, in all scenarios it is advantageous to use the study with highest quality for selecting candidate features at the beginning. In later steps we cannot compensate the effect of lost signals due to the application of a multiple testing adjustment on datasets with large noise.

Stepwise validation for simulated data with unequal sample sizes

As mentioned above, we assume that the quality of a dataset depends on the sample size(s) and the underlying noise. The effect of the latter was analyzed in the last paragraph, now our focus is on the influence of different sample sizes. We reduce the following descriptions to adjustment for multiple testing by controlling the FDR at a level of 5% because it seems sufficient at least for all parameter settings that we have

conducted.
The sample sizes of the three breast cancer datasets are used to generate simulated expression data. Hence, the simulated datasets have sample sizes of $n_{1M} = 47$ cases (metastasis) and $n_{0M} = 136$ controls (no evidence of metastasis) as in Mainz, $n_{1R} = 107$ cases and $n_{0R} = 168$ controls as in Rotterdam and $n_{1T} = 72$ cases and $n_{0T} = 189$ controls as in TRANSBIG. If we assume a different underlying noise in all datasets there are six possible arrangements for variance on datasets due to the different sample sizes. Analogously to the notation lq, mq, and hq for low, medium, and high quality we use the acronyms lv, mv, and hv for low, medium, and high variance, respectively. Tables 6 to 8 show the median of the true positives as well as the median of the false positives as before for the standard deviations 1.1, 1.5, and 2.3 for all possible sequences. We simulate a higher noise of 2.3 instead of 0.8, because in the previously described analyses one cohort (TRANSBIG) turned out to have a considerably higher noise than the other two studies. For example, $\sigma_M < \sigma_R < \sigma_T$ denotes that the variance of the data based on the sample sizes of the Mainz cohort is the lowest $\sigma_M^2 = 1.1$, the one based on the sample sizes of the Rotterdam study is medium $\sigma_R^2 = 1.5$, and the variance of the dataset with the sample sizes of TRANSBIG is the highest $\sigma_T^2 = 2.3$.

If we adjust the datasets with sample sizes of Rotterdam first and these datasets have low or medium variance, high numbers of true positives, 85 and 79, respectively, are obtained by the corresponding sequences after the third adjustment step. These numbers are only outperformed when the datasets with the sample sizes of TRANSBIG have the lowest underlying noise level and is adjusted first with 86 true positives. We obtain the worst results, i.e. smallest number of true positives, if a Mainz-like dataset is adjusted first. Especially, if this dataset has the highest underlying variance only three of the aimed 100 true positives were observed.

Summary

The simulation scenarios with equal sample sizes per group of patients suggest that it is the best strategy to start the sequential adjustment with the dataset of highest quality, i.e. lowest variance. Here, the highest number of true positives is obtained after three steps, independent of the order of the datasets used in the second and third step. Furthermore, as there are already no false positives after two adjustment steps, control-

	$\sigma_M < \sigma_R < \sigma_T$			$\sigma_M < \sigma_T < \sigma_R$		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $lv \to mv \to hv$	73 / 5	73.0 / 0	65 / 0	73 / 5	72 / 0	68.5 / 0
2: $lv \to hv \to mv$	73 / 5	64.0 / 0	64 / 0	73 / 5	69 / 0	68.0 / 0
3: $mv \to lv \to hv$	89 / 5	88.0 / 0	79 / 0	72 / 4	72 / 0	68.0 / 0
4: $mv \to hv \to lv$	89 / 5	79.0 / 0	79 / 0	72 / 4	68 / 0	68.0 / 0
5: $hv \to lv \to mv$	16 / 1	15.0 / 0	15 / 0	32 / 2	32 / 0	32.0 / 0
6: $hv \to mv \to lv$	16 / 1	15.5 / 0	15 / 0	32 / 2	32 / 0	32.0 / 0

Table 6: Median true positives/false positives in each step of the six possible adjustment sequences for setting $\sigma_M < \sigma_R < \sigma_T$ and $\sigma_M < \sigma_T < \sigma_R$ with low variance (lv) 1.1, medium variance (mv) 1.5 and high variance (hv) 2.3 (reproduced from Lohr et al., 2012).

	$\sigma_R < \sigma_M < \sigma_T$			$\sigma_R < \sigma_T < \sigma_M$		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $lv \to mv \to hv$	99 / 5	95 / 0	85 / 0	99 / 5	99 / 0	72 / 0
2: $lv \to hv \to mv$	99 / 5	88 / 0	85 / 0	99 / 5	71 / 0	71 / 0
3: $mv \to lv \to hv$	36 / 3	36 / 0	32 / 0	72 / 4	72 / 0	52 / 0
4: $mv \to hv \to lv$	36 / 3	32 / 0	32 / 0	72 / 4	52 / 0	52 / 0
5: $hv \to lv \to mv$	16 / 1	16 / 0	15 / 0	3 / 0	3 / 0	3 / 0
6: $hv \to mv \to lv$	16 / 1	15 / 0	15 / 0	3 / 0	3 / 0	3 / 0

Table 7: Median true positives/false positives in each step of the six possible adjustment sequences for setting $\sigma_R < \sigma_M < \sigma_T$ and $\sigma_R < \sigma_T < \sigma_M$ with low variance (lv) 1.1, medium variance (mv) 1.5 and high variance (hv) 2.3 (reproduced from Lohr et al., 2012).

	$\sigma_T < \sigma_M < \sigma_R$			$\sigma_T < \sigma_R < \sigma_M$		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $lv \to mv \to hv$	95 / 6	91 / 0	86 / 0	95 / 6	95 / 0	68.5 / 0
2: $lv \to hv \to mv$	95 / 6	90 / 0	86 / 0	95 / 6	68 / 0	68.0 / 0
3: $mv \to lv \to hv$	36 / 3	36 / 0	34 / 0	89 / 5	89 / 0	65.0 / 0
4: $mv \to hv \to lv$	36 / 3	34 / 0	34 / 0	89 / 5	64 / 0	64.0 / 0
5: $hv \to lv \to mv$	32 / 2	32 / 0	31 / 0	3 / 0	3 / 0	3.0 / 0
6: $hv \to mv \to lv$	32 / 2	31 / 0	31 / 0	3 / 0	3 / 0	3.0 / 0

Table 8: Median true positives/false positives in each step of the six possible adjustment sequences for setting $\sigma_T < \sigma_M < \sigma_R$ and $\sigma_T < \sigma_R < \sigma_M$ with low variance (lv) 1.1, medium variance (mv) 1.5 and high variance (hv) 2.3 (reproduced from Lohr et al., 2012).

ling the FDR seems sufficient. A third or even more adjustment steps are unnecessary in this scenario. Since we do not know the real underlying noise or the effect size we must be very careful about making generalizations from these conclusions.
The simulation study with unequal sample sizes suggests that first adjusting a dataset with largest sample size (as in the TRANSBIG cohort) and lowest variance yields the highest number of true positives. If the dataset with sample sizes of TRANSBIG is simulated with highest variance and the dataset with sample sizes of the Rotterdam cohort with lowest or at least medium variance, one should use the dataset with sample sizes of Rotterdam in the first adjustment step. Because we observe most significant probe sets when the real Rotterdam cohort is adjusted first, one may conclude that TRANSBIG has higher variance than Rotterdam and that Rotterdam has the highest quality, if we take the sample size and the underlying noise into account. Since TRANSBIG has no significant probe set when it is adjusted first, we may infer that is has lowest quality. Whether this depends on a high underlying variance or a different composition of the patient population remains unclear just as which of the datasets has the lowest or medium variance. Thus we can say that the 45 probe sets in the real breast cancer datasets, that are found to be significant after the second adjustment step, are very likely true positives and are therefore of biological interest.

3.2.3 Application to non-small cell lung cancer datasets

Next we apply the sequential validation algorithm to the non-small cell lung cancer datasets. Since the lung cancer datasets are hybridized on two different Affymetrix microarrays we take the overlap of 22 277 probe sets for further analyses. Instead of dividing the patients strictly in two distinct groups and looking for differential expressed genes or probe sets, we examine the correlation of expression values and the overall survival time of the patients. Therefore univariate Cox proportional hazards models are fitted to the expression values of every probe set in every dataset. We test the hypothesis "$H_0 : \mathrm{HR} = 1$" which is equivalent to "$H_0 : \beta = 0$" versus "$H_1 : \mathrm{HR} \neq 1$" with a Wald test. Altogether nine non-small cell lung cancer datasets are available. Thus, we have $9 \cdot 22\,277 = 200\,493$ p-values. In Table 9 the numbers of probe sets that are significant at a local α-level of 5% and 1% and on the same significance levels controlling the FDR for every single dataset are shown. The distributions of the unadjusted p-values can be

found in Figure 19 in the appendix. We see that the number of significant probe sets at an α-level of 5% for most datasets lies between 1 000 and 2 000. Exceptions are the Jacob dataset and GSE31210 that contain clearly more significant probe sets as well as GSE14814 that contains only 638 significant features at an α-level of 5%. After adjustment by controlling the FDR we find at least one significant probe set at an α-level of 5% (or 1%) only in the Jacob dataset, GSE31210, and GSE29013.

dataset	unadjusted 5%	unadjusted 1%	FDR 5%	FDR 1%
Uppsala	1875	450	0	0
Jacob	3402	1354	258	20
GSE4573	1118	189	0	0
GSE31547	1656	318	0	0
GSE3141	1492	366	0	0
GSE29013	1564	419	2	1
GSE31210	7597	4390	4355	1381
GSE19188	1177	190	0	0
GSE14814	638	128	0	0

Table 9: Numbers of significant probe sets unadjusted and after adjustment using FDR at a α-level of 5% and 1%, respectively, for all nine non-small cell lung cancer datasets.

In the last section two adjustment steps seemed sufficient to eliminate all false positive features. We apply one additional adjustment step, now then three adjustment steps, to the lung cancer datasets, because the power properties of the local Wald test might differ from the t-test and we might have missed the true underlying noise or the effect size in the real lung cancer datasets in our simulations. Thus $\frac{9!}{(9-3)!} = 504$ orders of three datasets out of the nine are possible. However, it makes less sense to consider sequences where one of the datasets that contain no significant features after simple adjustment controlling the FDR is set on first position. In fact, we merely find 13 sequences that yield at least one significant probe set after three adjustment steps. The numbers of significant probe sets in each step for those validation sequences are shown in Table 10. We see the datasets Jacob and GSE31210 as well as GSE31547 (on the third position) are rife in the list. For the sequences "Jacob $\rightarrow$ GSE31210 $\rightarrow$ GSE31547" and "GSE31210 $\rightarrow$ Jacob $\rightarrow$ GSE31547" the highest numbers of significant features are observed, 18 and 21, respectively. The six possible sequences for these three datasets and the correspond-

ing number of significant probe sets are listed in Table 23 in the appendix. In contrast GSE4573 and GSE14814 do not occur in any relevant sequence. GSE4573 consists of patients with the histological subtype of squamous cell carcinomas and GSE14814 contains almost twice as much squamous cell carcinomas as adenocarcinomas, while all patients in the Jacob dataset, GSE31547, and GSE31210 are diagnosed as adenocarcinomas. We may hypothesize that the squamous cell carcinomas biased the results in some way.

sequence					1st step	2nd step	3rd step
Jacob	→	Uppsala	→	GSE31547	258	5	1
Jacob	→	Uppsala	→	GSE31210	258	5	2
Jacob	→	Uppsala	→	GSE19188	258	5	3
Jacob	→	GSE29013	→	GSE31210	258	2	2
Jacob	→	GSE31210	→	Uppsala	258	120	2
Jacob	→	GSE31210	→	GSE31547	258	120	18
Jacob	→	GSE31210	→	GSE29013	258	120	2
GSE29013	→	GSE31547	→	Jacob	2	1	1
GSE31210	→	Jacob	→	Uppsala	4355	241	2
GSE31210	→	Jacob	→	GSE31547	4355	241	21
GSE31210	→	Jacob	→	GSE29013	4355	241	4
GSE31210	→	GSE29013	→	Jacob	4355	9	7
GSE31210	→	GSE29013	→	GSE3141	4355	9	2

Table 10: Numbers of significant probe sets in each step for the validation sequences (using FDR for significance adjustment and a α-level of 5%) that consider at least one significant probe set after three adjustment steps on the non-small cell lung cancer datasets.

For that reason the analyses are repeated restricted on the subgroup of adenocarcinomas. Eight datasets contain patients with adenocarcinomas and remain in the analysis, therefore $\frac{8!}{(8-3)!} = 336$ combinations of the datasets with respect to the order are possible. All orders are analyzed, but only six sequences yield at least one significant probe set after the third adjustment step. The six sequences and the numbers of significant features in each step regarding these sequences are listed in Table 24 in the appendix. The sequences "Jacob → GSE31210 → GSE31547" and "GSE31210 → Jacob → GSE31547" that reached the highest numbers of significant probe sets through all non-small cell lung cancer patients are in the list of relevant sequences for the adenocarcinomas, again, as we expected because all three datasets only contain patients with diagnosis adenocarci-

noma. In the four remaining sequences the dataset GSE3141 is represented at second or third position. If we have a look at the histogramms of the p-values of the Wald tests in the Cox proportional hazards models for all probe sets just containing the patients with an adenocarcinoma (see Figure 20 in the appendix), the distribution of those in GSE3141 has considerably changed compared to the distribution of all lung cancer patients in this dataset. The number of significant probe sets at a local α-level of 5% has increased to 2 722. The Uppsala dataset and GSE29013 are not contained in the list of sequences that lead to at least one significant feature after the third adjustment step of the algorithm.

3.3 Two-step meta-analysis approach

As mentioned at the beginning of the chapter the common approach to evaluate findings on several datasets is to apply meta-analyses to the effects of interest. In Botling et al. (2013) the meta-analysis approach is combined with the validation idea. In contrast to the sequential validation strategy above that uses at most three datasets in practise, all cohorts should be used in the following two-step approach. Since the Uppsala cohort is our basic dataset we start the procedure with this dataset. If we have a look on the p-values of the univariate Cox proportional hazards models in the Uppsala cohort as they were calculated in Section 3.2.3, again, no probe set holds a False Discovery Rate of 5% (cf. Table 9). This requires less strict adjustment on the Uppsala dataset to receive any potentially interesting probe set. Thus, all probe sets that pass an unadjusted significance-level of $p < 0.01$ in this first step are selected as possibly relevant genes. Again, like in the sequential approach in Section 3.2, all other features are ignored in the following second step. For the remaining 450 probe sets, meta-analyses are performed on the eight lung cancer datasets excluding the Uppsala cohort. Assuming a random effects model 59 probe sets show a raw p-value less than 0.01 and 13 p-values hold a FDR of 1%. We choose the random effects model instead of a fixed effects model, because we want to avoid testing for heterogenity for every probe set. Since eight datasets are available for the meta-analyses we can estimate an additional parameter, namely the between study effect. The workflow is illustrated in Figure 5.

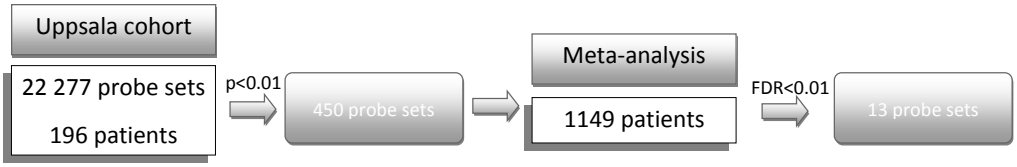

Figure 5: Workflow of the two-step meta-analysis (reproduced from Botling et al., 2013).

If we restrict to the histological subtype of adenocarcinomas, 658 probe sets are identified in the Uppsala cohort to be possibly interesting (raw p-value < 0.01). In the meta-

analysis less evidence for the significance of these features is found and only 32 probe sets yield an unadjusted p-value in the random effects model $p < 0.01$ and merely seven if the FDR of 1% is controlled.
Since dataset GSE31210 has most significant features of all cohorts and adjusting it first leads to the highest number of significant probe sets after three adjustment steps (cf. Section 3.2.3), we perform the analysis once more with GSE31210 as screening dataset. Like before we take the 4390 probe sets whose p-values in the univariate Cox proportional hazards model that pass the local α-level of 1% (cf. Table 9) as candidates to validate in the next step. In the meta-analysis 426 of the 4390 probe sets show a raw p-value < 0.01 assuming a random effects model and 63 hold a FDR of 1%. For the sake of completeness we restrict the analysis to the histological subgroup of adenocarcinomas once more. Due to the fact that all patients of dataset GSE31210 are diagnosed as adenocarcinomas, the first step remaines unchanged. Only the meta-analyses in the second step has to be restricted on the histological subgroup of adenocarcinomas on the other datasets. Then we observe 371 probe sets with a raw p-value less than 0.01 and 52 significant ones after adjustment with the Benjamini-Hochberg procedure with a FDR of 1%.

In Figure 6 we see the overlap of the significant probe sets at the end of the two-step

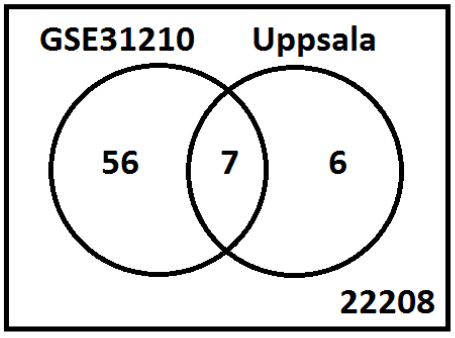

Figure 6: Visualisation of the significant features at the end of the two-step meta-analysis for all patients comparing the proceeding when Uppsala and GSE31210 are used as basic dataset, respectively.

meta-analysis when Uppsala is used as basic dataset compared with the results if dataset GSE31210 is used for preselection. The overlap is only seven probe sets that are listed with their corresponding gene symbols and gene names in Table 25 in the appendix.

3.4 Summary and comparison

Since microarray technology has become comparatively low priced, more and more gene expression data is freely available in appropriate databases. This information can help for validation of findings. Coinstantaneous, due to the large number of measured genes and possibly resulting statistical tests, it is necessary to control the probability for false positive findings. But after a strict adjustment for multiple testing and a subsequent just as stringent adjustment on every dataset, no significant results will be left. A common approach to combine the information of several datasets is the ordinary meta-analysis. However, this approach does not take the validation idea into account, but treats all datasets equally due to their variance. Hence, strategies that consider the validation idea and that are less strict in adjustment for multiple testing are required. In this chapter we presented two new approaches that meet the required demands.
The first approach was a sequential validation strategy that was proposed by Lohr et al. (2012). Here, an order of the considered datasets must be determined. The FDR or the FWER is controlled on the first dataset and only those features that hold the given α-level after adjusting the p-values by the Benjamini-Hochberg or rather Bonferroni-Holm procedure are examined on the next cohort. We applied this procedure to the Mainz, Rotterdam and TRANSBIG breast cancer datasets, performed extensive simulation studies and adopt the strategy to nine non-small cell lung cancer cohorts at last. A finding of the simulation study was that two adjustment steps and the control of the FDR seemed sufficient to eliminate all false positive features in the conducted settings. For the reliability of non-false positive results a conservative additional step was carried out on the lung cancer datasets. However, the application of this 3-step strategy remained challenging, since merely three datasets, namely Jacob, GSE31210, and GSE29013, yield at least one probe set that is significant after adjustment by controlling the FDR on a level of 5%. The sequences "Jacob $\rightarrow$ GSE31210 $\rightarrow$ GSE31547" and "GSE31210 $\rightarrow$ Jacob $\rightarrow$ GSE31547" render the highest numbers of significant features (18 and 21, respectively) after the third step. The top results remain the same if the analysis is restricted to patients with the histological subtype of adenocarcinomas, since these three cohorts consist only of patients that are diagnosed as adenocarcinomas. If we consider all patients eleven more 3-step sequences lead to at least one significant probe set, in the subgroup of adenocarcinomas only four sequences. But the numbers of significant probe sets after three steps are generally higher in the mentioned sequences

when we restrict to the adenocarcinomas instead of analyzing all patients, although the number of patients decreases. On the other hand particularly the Uppsala cohort seems to have less impact if constrained to the adenocarcinomas in comparison to all patients. That points to a common issue of subgroup analysis. If we want to draw any conclusion about a subgroup and use another, hopefully similar, subgroup besides the group of interest, the results will be biased. If the patients of other subgroups are ignored we have a smaller sample size and the variance will increase. For example Netzer (2013) has a closer look on the bias-variance-tradeoff in subgroup analysis. Here, all patients are considered for the analysis of a special subgroup but depending on their similarity/characteristics with different weights.

The greatest advance in the sequential valididation approach is that the quality of a dataset, that we define by sample size and the underlying noise in the data, may be assessed compared to other datasets in some situations. If we have a look on the results of the breast cancer datasets in Section 3.2.1, again, it is obvious that TRANSBIG will have the highest underlying noise, because there are more patients in this study than in the Mainz cohort and adjusting TRANSBIG (first) we find no significant features. We may conjecture that it is caused by a batch effect because this cohort consists of patients data from five different European centers. Rotterdam is the dataset with highest quality in terms of sample size and variance. But whether Mainz or Rotterdam has less underlying noise remains unclear since more patients are included in the Rotterdam cohort. Another important issue of the sequential validation approach remains unsolved.

The simulation study pointed out that two adjustment steps are sufficient to eliminate all false positive features, which was our priority objective. However, in simulation studies it is impossible to regard all scenarios since we do not know the true number of significant features, the true effect size, and so on. In addition, all settings cannot be repeated for every testing procedure. Therefore, we applied an additional, third adjustment step to the real lung cancer datasets. Yet, it is questionable that three adjustment steps are the optimal strategy in either case. Three steps may be not sufficient to eliminate all false positives, e.g. if the number of true positives might be exceeding high or the effect size is quite small. In other situations, as in the simulations, two steps are sufficient and through a third step the power of the procedure decreases. A better approach will probably be to take all available datasets for the validation of results into consideration. In Section 3.3 a workflow that includes all given datasets is introduced. This approach does not require a selection among the available datasets nor an assessment of the com-

plete order for the validation. The procedure is simple. A preselection of potentially interesting candidate features is made without any adjustment on the basic dataset and afterwards the effects of these genes are validated via meta-analyses on the remaining cohorts. The adjustment in the first step is missing because as few as possible true positive features shall be overlooked. Pursuing this workflow, we yield 450 probe sets whose p-values hold a local α-level of 1% in the Uppsala dataset. Performing meta-analyses on the other eight non-small cell lung cancer datasets for these 450 preselected probe sets yields 13 features that are significant even when controlling the FDR (59 with a raw p-value < 0.01). The histological subgroup of patients with adenocarcinomas were analysed with this approach, too. Here, 658 interesting probe sets with a raw p-value < 0.01 are found in the Uppsala cohort, which is nearly one and a half times as much as in all non-small cell lung cancer patients, although just half the amount of patients are considered. Applying meta-analyses to the adenocarcinoma patients of the remaing seven datasets that contain patients of this histological subtype to these 658 features only yields seven probe sets whose p-values hold a FDR of 1% (32 with a raw p-value < 0.01). It seems as there are less signals in the subgroup of adenocarcinoma patients than in the other datasets which might be up to the smaller sample size but also to quality criteria as e.g. minor fraction of tumour cells in the samples or inaccuracy among the histological grading. Since cohort GSE31210 brings up most significant probe sets (4355 hold a FDR of 5% and 4390 a local α-level of 1%) we repeat the procedure with GSE31210 as basic dataset. Following the workflow we yield 63 probe sets that hold a FDR of 1% (426 probe sets with corresponding p-values $<$ local α-level of 0.01) if all non-small cell lung cancer patients are considered and 52 probe sets that hold a FDR of 1% (371 probe sets $<$ local α-level of 0.01) after all if we restrict to the subgroup of adenocarcinoma patients. We see that even in this approach it is crucial which dataset is considered first. If we assume the dataset with the highest number of significant features to have highest quality, it would be beneficial to start with even this cohort to obtain most (true) significant genes.

As a comparision with the two validation approaches we perform a common meta-analysis that ignores the validation idea. A random effect model is assumed for each probe set of the overlap of the two different Affymetrix arrays to avoid an additional test for heterogenity. Analyzing all non-small lung cancer patients 123 of the 22 277 probe sets hold a FDR of 1% (1665 raw p-values < 0.01), while 120 probe sets are significant on a FDR of 1% if only the adenocarcinomas are considered. Most of the features that were

identified in the validation meta-analysis approach using Uppsala as well as GSE31210 as basic dataset were found with the common meta-analyses.
The probe set with the smallest p-value in the random effects model of all nine non-small cell lung cancer datasets that is also significant in the validation meta-analysis if Uppsala as well as if GSE31210 is taken as basic dataset represents gene "AGFG1" that encodes a protein that is related to nucleoporins that are responsible for mediating nucleocytoplasmic transport (Fritz et al., 1995). The corresponding forest plot is illustrated in Figure 7.

We see that the estimated Hazard Ratios of all studies are greater than 1 and the studies quite agree with each other, all confidence intervals comprise the overall study effect.

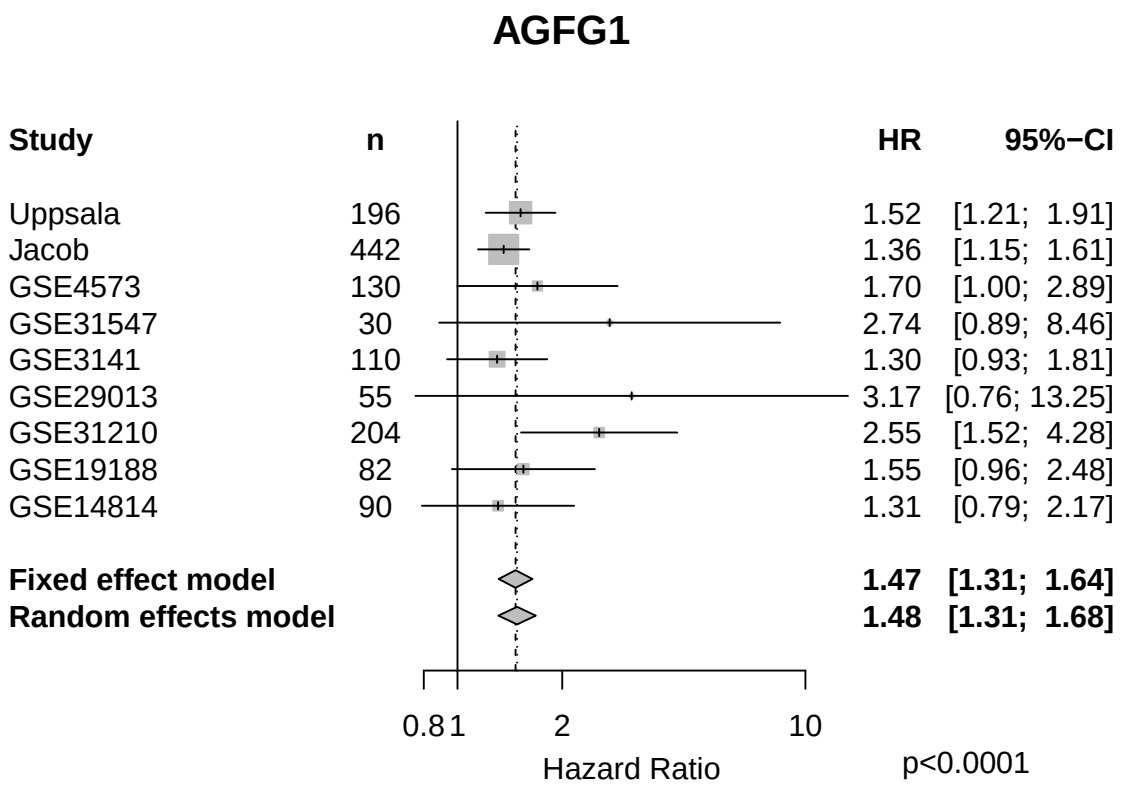

Figure 7: Forestplot of the meta-analysis for probe set "218092_s_at" that represents the gene "AGFG1" including all nine non-small cell lung cancer datasets. The p-value (bottom right) corresponds to the random effects model.

For this reason the fixed effect and the random effects model yield very similar results. Four of the studies are even significant considering the single study effect, particularly Uppsala and GSE31210. The estimated hazard ratios of the datasets with smallest sample sizes (GSE31457 and GSE29013) have the largest confidence intervals. Altogether the p-value < 0.0001 of the random effects model indicates that the true hazard ratio is unequal to zero. More precisely, if the expression of AGFG1 increases, the overall survival time of non-small lung cancer patients decreases. Thus, AGFG1 seems to be an oncogene.

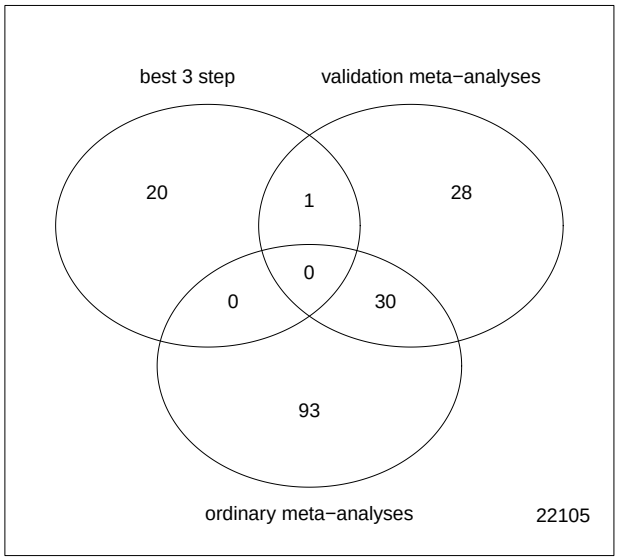

Figure 8: Visualisation of the significant features at the end of the best 3-step sequential validation approach, the combined meta-analysis for all patients comparing the proceeding when Uppsala is used as basic dataset, and the common meta-analyses assuming a random effects model.

The comparison of the results of the validation meta-analysis, the common meta-analysis, and the sequential validation approach is visualized in Figure 8. Since all 3-step sequences were considered in the latter we look at the results of the sequence that yields the highest number of significant features. In Figure 8 it is recognizable that the great-

est overlap of significant features can be found between common and validation meta-analyses, while there is no overlap between the significant probe sets of the ordinary meta-analyses and the best 3-step validation analysis, and therefore neither in the overlap of all three aproaches. The validation meta-analyses and the best 3-step validation share only one significant feature: The probe set "218451_at" represents the gene "CDCP1" which encodes the "CUB domain containing protein 1". From Brown et al. (2004) it is known that CDCP1 is overexpressed in carcinomas, and that CDCP1 mRNA is highly elevated in human lung cancer cells (Scherl-Mostageer et al., 2001). Ideka et al. (2006) even discovered that tumors with higher expression of CDCP1 show a higher level of proliferation than tumors with low CDCP1 expression. Thus, our results emphasize the known findings from literature.

In Figure 21 in the appendix the forest plot of the probe set "218451_at" that represents CDCP1 can be seen. Except for the estimated hazard ratio of study GSE3141 all single study effects are greater than 1. Five of the nine studies are significant if considered seperately. The combined study effect of the random effects model is 1.41 (CI: $[1.16 - 1.71]$) which leads to a p-value of 0.0005. Due to the high number of analyzed features that require adjustment for multiple testing the p-value does not hold a FDR of 5%.

Since the best 3-step validation sequence starts with GSE31210, it is more suitable to compare the results of this sequence with a validation meta-analysis where GSE31210 is used as basic dataset. The overlap of the significant features hereof and the ordinary meta-analyses are visualized in Figure 22 in the appendix. We find 426 significant probe sets with the validation meta-analysis when GSE31210 is screened at the beginning. Although the number of significant features increases with this proceeding the overlap with the results of the common meta-analyses is not considerably higher. Again, there is no overlap of the significant features between the validation meta-analyses and the ordinary meta-analyses. However, the validation meta-analyses and the best 3-step validation sequence have six significant probe sets in common. The list of significant probe sets including the genes that they are representing can be found in Table 11.

probe set	symbol	gene name
201251_at	PKM2	pyruvate kinase, muscle
201546_at	TRIP12	thyroid hormone receptor interactor 12
209313_at	GPN1	GPN-loop GTPase 1
218451_at	CDCP1	CUB domain containing protein 1
212581_x_at	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
AFFX-HUMGAPDH/ M33197_M_at	GAPDH	glyceraldehyde-3-phosphate dehydrogenase

Table 11: Probe sets with gene symbol and gene name that are significant after three steps of sequential validation of the order "GSE31210 → Jacob → GSE31547" as well as in the validation meta-analysis when dataset GSE31210 is used for preselection.

Like above the gene CDCP1 is in the overlap, again. It is conspicuous that two probe sets of the gene "GAPDH" that encodes the enzyme glyceraldehyde-3-phosphate dehydrogenase are in the list of the overlap. From literature it is known that GAPDH is overexpressed in lots of tumors (Sirover 1999, Said et al. 2009) and that overall survival and the relapse-free survival are reduced in patients whose level of GAPDH expression is enhanced (Révillion et al. 2000). These results can be confirmed with our findings (cf. Figures 23 and 24 in the appendix). The estimates for the hazard ratio in all studies are highly correlated for the two probe sets. We see a slightly smaller p-value for probe set "212581_x_at", because here the estimated hazard ratio of GSE19188 and GSE29013 points towards the right direction compared to probe set "M33197_M_at".
In conclusion, it can be claimed that the ordinary meta-analysis is an excellent method for the analysis of several datasets, most significant features that hold a FDR 1% can be found by the application of this procedure. But apart from neglecting the validation idea, important genes were not identified by the ordinary meta-analysis, e.g. CDCP1. The assessment of study quality with the stepwise validation approach described in Section 3.2 might be useful, but it remains difficult in most cases. If the number of significant features in combination with intrinsic noise are considered as quality criteria, the Rotterdam cohort seems to be of highest quality of the breast cancer studies. Among the non-small cell lung cancer datasets the Jacob cohort and GSE31210 seem to have the highest quality. If we take a closer look, it is obvious that the "good" datasets consist

of only one histological subtype. In addition to sample size and noise, something like biological comparability or rather consistency must be included in the constitution of quality. Therefore, the definition of quality remains challenging and should be reconsidered for every problem.

4 Differential gene expression networks

The reconstruction of biological, i.e. gene-gene, protein-protein or gene-protein, networks is a recent research topic (e.g. Juric et al., 2007; Gill et al., 2010). It is known that genes do not act independently, but groups of genes act and interact with each other. The estimation of gene regulatory networks is challenging.

The dependence structure of g genes is often of interest. Though we get an impression of the overall correlation structure by calculating the ordinary correlation coefficients ρ_{il}, $i, l = 1, \ldots, g$, the actual dependencies are not recognizable, because ordinary correlation coefficients do not distinguish between direct and indirect interactions. A direct interaction between two variables w.l.o.g. X_1 and X_2 occurs if X_1 has a direct influence on X_2 or vice versa. This is in contrast to an indirect interaction where X_1 and X_2 are e.g. both influenced by X_3 and conditioned on X_3 they become independent. Ordinary (estimated) correlations are therefore only weak evidence for real direct dependencies between two genes, while the absence of correlation argues for independence. Hence, an adaption is required that considers the dependence structures of other given variables. Precisely the partial correlation coefficient accounts for this issue. It quantifies the correlation between two variables X_1 and X_2 conditioning on serveral other variables, or in other words, it is defined as the correlation between the residuals of a linear regression explaining X_1 and X_2, respectively, with the other variables as covariates (Fujikoshi et al., 2010).

In the following we denote random variables by capital letters, the corresponding observations by small form letters. Matrices are printed bold, vectors are marked by an arrow " $\rightharpoonup$ " and estimators are labeled by " $\hat{}$ " above the letter. Figure 9 i) shows a direct interaction of A and B, where A is a parent of B. This means A has a direct influence on child node B. In Figure 9 ii) we see a directed path from B to A. But conditioned on C, A and B are independent, as well as in Figure 9 iii), where C has the children

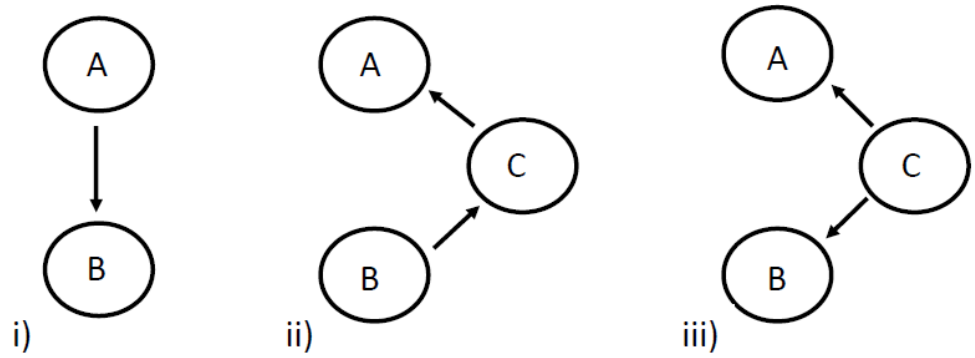

Figure 9: Examples for a i) direct interaction between A and B, ii) indirect interaction between A and B and iii) interaction of A and B by regulation by a common gene C.

A and B, and A and B are not adjacent. While for situations i) to iii) of Figure 9 the ordinary Pearson correlation coefficient will recognize a non-zero correlation between A and B, the partial correlation for the indirect interaction between A and B in ii) and the apparent interaction of A and B due to a common regulator C in iii) will 0. We see, the ordinary correlation is weak evidence for measuring dependence, since in our example more or less all gene pairs will have non-zero correlation. In contrast, partial correlations provide only a weak criterion for independence, since most partial correlation vanish, but it offers a strong measure of dependence.

In this thesis we go one step further. Our focus will not lie on the reconstruction of genetic networks, but it is on the detection of differential networks. Therefore, we introduce several methods for the identification of differential networks in the next section. Afterwards, we present some approaches for the detection of differential genetic networks and subsequently conduct an extensive simulation study on tests for recognizing differential networks. In Section 4.4 the results will be summarizied.

4.1 Methods for differential network analysis

A popular tool for the analysis of gene association networks are Graphical Gaussian Models (GGMs), also named Covariance Selection Models following Dempster (1972) who first suggested to fit models with zeros in the off-diagonal elements of the concentration matrix, i.e. the inverse of the covariance matrix. The basis of GGMs form so called partial correlations as measures of conditional independence. In contrast to relevance networks (see e.g. Butte et al., 2000) that use the standard pearson correlation coefficient and a predefined threshold, GGMs are able to distinguish between direct interactions between two genes, indirect interactions, and regulation by a third common gene.
In the following Section 4.1.1 the link between probability theory and graph theory is clarified. After the general definition of a graphical model, GGMs as graphical models under the assumption of a multivariate normal distribution are introduced. Partial correlations that form the basis of GGMs are introduced in Section 4.1.2. Since partial correlations require a reliable estimation of the covariance matrix a shrinkage approach for the latter is described in chapter 4.1.3. How a Graphical Gaussian Model is obtained by the use of a heuristic mixture model approach is explained in Section 4.1.4.
After the estimation of the GGMs, we go a step further and want to discover differences in gene association networks under two conditions. Hence, several measures for the comparison of two networks are introduced in Section 4.1.5. Since the developed measures follow no known probability distribution, permutation tests are required that are briefly reviewed in Section 4.1.6.

4.1.1 The link from probability theory to graph theory

A *graph* $\mathcal{G}$ is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of a finite set of *nodes* (or *vertices*) $\mathcal{V}$ and a finite set of *edges* $\mathcal{E}$ between the edges in $\mathcal{V}$, i.e. $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ (Edwards, 2000). An *undirected edge* between nodes $\delta, \gamma \in \mathcal{V}$ can therefore be written as (δ, γ) or (γ, δ) and it is visualized by a line. If all edges are undirected we call such a graph *undirected*. In contrast, if we consider a graph $\mathcal{G}^{\text{`}} = (\mathcal{V}^{\text{`}}, \mathcal{E}^{\text{`}})$, where the edges are directed, i.e. if $(\gamma, \delta) \in \mathcal{E}$ but $(\delta, \gamma) \notin \mathcal{E}$, we call it *directed*. Since we focus on undirected graphs in this thesis, we restrict to those in the introduction of notation, terminology, and properties in the following.

A *path* is defined as a sequence of edges (not necessarily directed) that is connected (Lauritzen, 1996). A *subgraph* $\mathcal{G}_A$ is a graph restricted to a subset of nodes $A \subseteq \mathcal{V}$ and a subset of vertices $\mathcal{E}_A = \mathcal{E} \cap (A \times A)$. We call a graph *complete* if all nodes are connected by edges. If there is an edge between δ and γ, they are said to be *neighbours* or *adjacent* and δ and γ are called *non-adjacent*, if no line between the two vertices exists. The set of vertices that are connected to δ is denoted by $ne(\delta)$. For a subset $A \subseteq \mathcal{V}$ the collection of neighbours of nodes in A is defined as the union of neighbours that are not in A, $ne(A) = \bigcup_{\delta \in A} ne(\delta) \setminus A$. If we speak of undirected graphs the set of neighbours $ne(A)$ is also referred to as *boundary* of A. The union of A and $ne(a)$ is called *closure*, $cl(A) = ne(A) \cup A$. A subset $B \subseteq \mathcal{V}$ *separates* two vertices δ and γ if all paths from δ to γ intersect B. For $A, B, C \subseteq \mathcal{V}$ we say that B *separates* A *from* C, if B separates all nodes $\delta \in A$ from $\gamma \in C$.
A toy example for illustration is given in Figure 10. We define the subsets $A \equiv \{X_1, X_2, X_3\}$, $B \equiv \{X_4\}$, and $C \equiv \{X_5, X_6\}$. Then the collections of neighbours for the three distinguished subsets are $ne(A) = \{X_4\}$, $ne(B) = ne(\{X_4\}) = \{X_2, X_3, X_5, X_6\}$, and $ne(C) = \{X_4\}$, while the closure of A is given by $cl(A) = \{X_1, X_2, X_3, X_4\}$, the closure of B by $cl(B) = \{X_2, X_3, X_4, X_5, X_6\}$, and analogous the closure of C by $cl(C) = \{X_4, X_5, X_6\}$. B or rather X_4 separates $A = \{X_1, X_2, X_3\}$ from $C = \{X_5, X_6\}$ because all paths from nodes of A to those of C intersect X_4.

The idea for the link between graph theory and probability theory is to use a graph for the illustration of an association structure of random variables. Thereby, nodes stand for the random variables and edges represent (conditional) independence structures among the variables (Lauritzen, 1996). Hence, we clarify the concept of conditional independence first.
Let X_1, X_2, and X_3 be random variables with a joint distribution P. Then X_1 *is conditionally independent of* X_2 *given* X_3 *under* P if for any measurable set A_1 in the sample space of X_1 $P(A_1|X_2, X_3) = P(A_1|X_2)$, i.e. the conditional probability of A_1 given X_2 and X_3 is independent of X_3 and we write $X_1 \perp X_2|X_3\,[P]$, or short $X_1 \perp X_2|X_3$. If the random variables have a continuous density it holds that

$$X_1 \perp X_2|X_3 \Leftrightarrow f_{X_1X_2X_3}(x_1, x_2, x_3)\, f_{X_3}(x_3) = f_{X_1X_3}(x_1, x_3)\, f_{X_2X_3}(x_2, x_3)\,.$$

Let $U = h(X_1)$ be an arbitrary measurable function on the sample space of X_1 and X_4

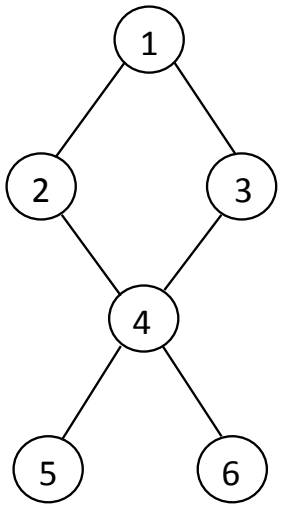

Figure 10: Toy example.

another random variable. And let X_1 be conditional independent of X_2 given X_3, i.e. $X_1 \perp X_2 | X_3$. An algebraic structure that satisfies

(G1) X_2 is conditionally independent of X_1 given X_3, i.e. $X_2 \perp X_1 | X_3$;

(G2) U is conditionally independent of X_2 given X_3, i.e. $U \perp X_2 | X_3$;

(G5) X_1 is conditionally independent of X_2 given X_3 and U, i.e. $X_1 \perp X_2 | (X_3, U)$;

(G4) If in addition X_1 is conditionally independent of X_4 given X_2 and X_3, then X_1 is conditionally independent of X_2 and X_4 given X_3, i.e. $X_1 \perp (X_2, X_4) | X_3$;

where $U \subseteq X_1$ and X_1, X_2, X_3 are disjoint and finite subsets, is called *semi-graphoid*. If X_1, X_2, X_3 are disjoint and the joint distribution of all variables is positive and continuous

(G5) If in addition X_1 is conditionally independent of X_3 given X_2, then X_1 is conditionally independent of X_2 and X_3, i.e. $X_1 \perp (X_2, X_3)$.

will hold and the algebraic structure is called *graphoid*.

The concept of graph separation is an example of an algebraic structure that fullfilles the semi-graphoid axioms, so

$$A \overset{\mathcal{G}}{\perp} C|B \qquad \Leftrightarrow \qquad B \,\text{separates}\, A \,\text{from}\, C.$$

If the subsets A, B, and C are disjoint the algebraic structure satisfies the graphoid axioms $(G1)$-$(G5)$. Coming back to the toy graph in Figure 10 we see that A is independent of C given B because $B = \{X_4\}$ separates $A = \{X_1, X_2, X_3\}$ from $C = \{X_5, X_6\}$.

For an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a collection of random variables $(X_\delta)_{\delta \in \mathcal{V}}$ that take values into probability space three so called *Markov properties* are defined as follows:

(P) All non-adjacent pairs of vertices are independent conditional on the remaining nodes: If for a given graph and all non-adjacent nodes, $\delta, \gamma \in \mathcal{V}$, it holds that

$$\delta \perp \gamma \,|\, \mathcal{V} \backslash \{\delta, \gamma\}\,,$$

the probability measure is said to obey the *pairwise Markov property.*

(L) Conditional on the adjacent vertices, any vertex is independent of the remaining nodes: A probability measure satisfies the *local Markov property* if for any vertex $\delta \in \mathcal{V}$, δ is independent of $\mathcal{V}$ with the closure of δ given the boundary of δ, i.e.

$$\delta \perp \mathcal{V} \backslash cl\,(\delta) \;|\, bd\,(\delta)\,.$$

(G) Any two disjoint subsets of nodes separated by a third subset is conditionally independent given the vertices in the third subset: The *global Markov property* is fulfilled by a probability measure if B separates A from C in $\mathcal{G}$ for all disjoint subsets $A, B, C \in \mathcal{V}$, i.e.

$$A \overset{\mathcal{G}}{\perp} C|B.$$

The global Markov property implies the local, while the local implies the pairwise Markov property. If property $(G5)$ of the graphoid axioms is satisfied all three Markov properties are equivalent, and therefore graph separation satisfies the graph axioms.

Since conditional independence is highly related to factorization, so are the Markov properties. Let $A \subseteq \mathcal{V}$ be a complete subset of $\mathcal{G}$. We say a probability distribution P *factorizes* according to $\mathcal{G}$ if for all such A exists a non-negative function ψ_A that depends on the A-coordinates alone, and P has a density f that factorizes to

$$f(x) = \prod_{A\,\text{complete}} \psi_A(x_A),$$

where $x = (x_v,\, v \in V)$ and $x_A = (x_v,\, v \in A)$. In general, it holds that for every undirected graph the factorization property implies the global Markov property. *Hammersley and Clifford* (1971; Clifford, 1990) showed that if a probability distribution satisfies the pairwise Markov property and has a positive and continuous density if and only if it factorizes according to $\mathcal{G}$.

An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a set of random variables that satisfies the local Markov property is called a *Markov network* or *graphical model* (Kindermann and Snell, 1980).
As a result of the theorem of Hammersley and Clifford it is sufficient to show the pairwise Markov property, if the probability distribution has a positive and continuous density, to proof the required local Markov property for graphical models. A special case of graphical models are so called *Covariance Selection Models* or *Graphical Gaussian Models* (GGMs). As the name suggests the underlying probability distribution of the random variables is a multivariate normal (Gaussian) distribution. Let $\vec{X}$ be a random vector and assume it to be multivariate normal distributed with $\vec{X} \sim N_p\left(\vec{\mu}, \mathbf{\Sigma}\right)$, where $\mathbf{\Sigma}$ is assumed to be regular. W.l.o.g. partition $\vec{X}$ in $\vec{X} = (X_1, X_2, \vec{X}_3)'$, where X_1 and X_2 are random variables and $\vec{X}_3 = (X_3, \ldots, X_p)'$ denotes a vector of random variables. In Lauritzen (1996, Proposition 5.2, p.129) it is shown that

$$X_1 \perp X_2 | X_3 \qquad \Leftrightarrow \qquad \omega_{12} = 0,$$

where $\Omega = \{\omega_{il}\}_{i,l \in \{1,\ldots,p\}} = \Sigma^{-1}$ is the concentration matrix of the multivariate normal distribution. We see, the multivariate normal distribution is positive and continuous and it obeys the pairwise Markov property and thereby also the local and global Markov properties and the factorization property.

In the next section, partial correlations that form the basis of Graphical Gaussian Models due to the association with the concentration matrix are derived.

4.1.2 Partial correlations

Let w.l.o.g. X_1 and X_2 be two random variables, $\vec{X}_3 = (X_3, \ldots, X_p)^{\prime}$ denotes a vector of random variables, and $\vec{X} = (X_1, X_2, \vec{X}_3)^{\prime}$ with corresponding mean vector $\vec{\mu}$ and covariance matrix $\mathbf{\Sigma}$, where

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vec{\mu}_3 \end{pmatrix} \qquad \text{and} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \vec{\sigma}_{31}^{\prime} \\ \sigma_{21} & \sigma_{22} & \vec{\sigma}_{32}^{\prime} \\ \vec{\sigma}_{31} & \vec{\sigma}_{32} & \mathbf{\Sigma}_{33} \end{pmatrix}.$$

The n realizations of the random variables or rather vectors are given by $\vec{x}_1$, $\vec{x}_2$, and $\boldsymbol{x}_3$, respectively, with

$$\vec{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \end{pmatrix}, \qquad \vec{x}_2 = \begin{pmatrix} x_{21} \\ \vdots \\ x_{2n} \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{x}_3 = \begin{pmatrix} x_{31} & \cdots & x_{p1} \\ \vdots & & \vdots \\ x_{3n} & \cdots & x_{pn} \end{pmatrix}.$$

As mentioned above partial correlation can be described by the correlation of the residuals of linear models where the variables of interest (X_1 and X_2) are explained by the remaining variables that are considered ($\vec{X}_3$). In general, the best linear predictor can by obtained by the least squares estimator that is given by

$$\hat{y} = \hat{\beta}\boldsymbol{z},$$

where $\hat{\beta} = (\beta_1, \ldots, \beta_p)^{\prime}$ is the least squares parameter vector and $\boldsymbol{z}$ is the design matrix (cf. Toutenburg, 2009). To derive the best linear predictors for X_1 and X_2, we define $\vec{x}_i^{\star} := \vec{x}_i - \bar{x}_i \in \mathbb{R}^n$, $i = 1, 2$, $\boldsymbol{x}_3^{\star} := \boldsymbol{x}_3 - \bar{x}_3 \in \mathbb{R}^{n \times (p-2)}$, where $\bar{x}_i$, $i = 1, 2$, are the sample means of $\vec{x}_1$ and $\vec{x}_2$, respectively, and $\bar{x}_3 = (\bar{x}_{33}, \ldots, \bar{x}_{3p})^{\prime}$ is a sample mean vector, as

transformations of the observations and

$$\hat{x}_i^\star := \hat{x}_i - \bar{x}_i = \begin{pmatrix} \hat{x}_{i1} \\ \vdots \\ \hat{x}_{in} \end{pmatrix} - \bar{x}_i \in \mathbb{R}^n,\ i = 1, 2$$

as transformations of the estimated values.
We look at the predictor for the transformed values

$$\hat{x}_i^\star = \boldsymbol{x}_3^\star \cdot \hat{\beta}_i. \tag{1}$$

The general least squares estimator for β_i is given by

$$\hat{\beta}_i = (\boldsymbol{x}_3^{\star\text{`}} \boldsymbol{x}_3^\star)^{-1} \boldsymbol{x}_3^{\star\text{`}} \cdot \vec{x}_i^\star,\ i = 1, 2.$$

This formula can be transformed to

$$\begin{aligned}
\hat{\beta}_i &= (\boldsymbol{x}_3^{\star\text{`}} \boldsymbol{x}_3^\star)^{-1} \boldsymbol{x}_3^{\star\text{`}} \cdot \vec{x}_i^\star \\
&= \left[(\boldsymbol{x}_3 - \bar{x}_3)^\text{`} (\boldsymbol{x}_3 - \bar{x}_3) \right]^{-1} \cdot (\boldsymbol{x}_3 - \bar{x}_3)^\text{`} \left(\vec{x}_i - \bar{x}_i \right) \\
&= \underbrace{\left[\frac{1}{n} (\boldsymbol{x}_3 - \bar{x}_3)^\text{`} (\boldsymbol{x}_3 - \bar{x}_3) \right]^{-1}}_{\hat{\Sigma}_{33}^{-1}} \cdot \underbrace{\left[ \frac{1}{n} (\boldsymbol{x}_3 - \bar{x}_3)^\text{`} \left(\vec{x}_i - \bar{x}_i \right) \right]}_{\hat{\sigma}_{3i}} \\
&= \hat{\Sigma}_{33}^{-1} \cdot \hat{\sigma}_{3i},
\end{aligned}$$

where $\hat{\sigma}_{i3}$ and $\hat{\Sigma}_{33}^{-1}$ are estimates of σ_{i3} and Σ_{33}^{-1}, respectively. If we insert this in expression (1), we yield

$$\begin{aligned}
& & \hat{x}_i^\star &= \boldsymbol{x}_3^\star \cdot \hat{\Sigma}_{33}^{-1} \cdot \hat{\sigma}_{3i} \\
\Leftrightarrow & & \vec{x}_i - \bar{x}_i &= (\boldsymbol{x}_3 - \bar{x}_3) \cdot \hat{\Sigma}_{33}^{-1} \cdot \hat{\sigma}_{3i} \\
\Leftrightarrow & & \vec{x}_i &= \bar{x}_i + (\boldsymbol{x}_3 - \bar{x}_3) \cdot \hat{\Sigma}_{33}^{-1} \cdot \hat{\sigma}_{3i} \\
\Leftrightarrow & & \vec{x}_i &= \bar{x}_i + \left(\hat{\sigma}_{3i}^{\text{`}} \cdot \hat{\Sigma}_{33}^{-1} \cdot (\boldsymbol{x}_3 - \bar{x}_3)^\text{`} \right)^\text{`}
\end{aligned}$$

Because $\left(\vec{\sigma}_{i3}^{\,\prime} \cdot \boldsymbol{\Sigma}_{33}^{-1} \cdot \left(\vec{X}_3 - \vec{\mu}_3\right)\right)^{\prime} = \vec{\sigma}_{i3}^{\,\prime} \cdot \boldsymbol{\Sigma}_{33}^{-1} \cdot \left(\vec{X}_3 - \vec{\mu}_3\right) \in \mathbb{R}$, the best linear predictor of X_i, $i = 1, 2$, by a linear function of $\vec{X}_3$ is

$$l_i(\vec{X}_3) := \hat{X}_i = \mu_i + \vec{\sigma}_{i3}^{\,\prime} \cdot \boldsymbol{\Sigma}_{33}^{-1} \cdot \left(\vec{X}_3 - \vec{\mu}_3\right).$$

The residuals

$$X_i^{\star} := X_i - l_i(\vec{X}_3),\ i = 1, 2,$$

are defined as the remaining portion of X_i, $i = 1, 2$, after removing the linear effects of X_3 and we define the vector of residuals as

$$\vec{X}^{\star} := \begin{pmatrix} X_1^{\star} \\ X_2^{\star} \end{pmatrix} = \begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} - \begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1} \left(\vec{X}_3 - \vec{\mu}_3\right),$$

with

$$\begin{aligned}
\operatorname{Var}\left(\vec{X}^{\star}\right) &= \operatorname{Var}\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} - \begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1} \left(\vec{X}_3 - \vec{\mu}_3\right)\right] \\
&= \operatorname{Var}\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix}\right] + \operatorname{Var}\left[\begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1} \left(\vec{X}_3 - \vec{\mu}_3\right)\right] \\
&\quad - 2 \cdot \operatorname{Cov}\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix}, \begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1} \left(\vec{X}_3 - \vec{\mu}_3\right)\right] \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} + \begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1} \boldsymbol{\Sigma}_{33} \left[\begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \boldsymbol{\Sigma}_{33}^{-1}\right]^{\prime} \\
&\quad - 2 \cdot \left[\begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \left(\boldsymbol{\Sigma}_{33}^{-1}\right)^{\prime} \begin{pmatrix} \vec{\sigma}_{31} & \vec{\sigma}_{32} \end{pmatrix}\right] \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} + \begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \left(\boldsymbol{\Sigma}_{33}^{-1}\right)^{\prime} \begin{pmatrix} \vec{\sigma}_{31} & \vec{\sigma}_{32} \end{pmatrix} \\
&\quad - 2 \cdot \left[\begin{pmatrix} \vec{\sigma}_{31}^{\,\prime} \\ \vec{\sigma}_{32}^{\,\prime} \end{pmatrix} \left(\boldsymbol{\Sigma}_{33}^{-1}\right)^{\prime} \begin{pmatrix} \vec{\sigma}_{31} & \vec{\sigma}_{32} \end{pmatrix}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \overrightarrow{\sigma}_{31}^{\,\prime} \\ \overrightarrow{\sigma}_{32}^{\,\prime} \end{pmatrix} \left(\mathbf{\Sigma}_{33}^{-1}\right)^{\prime} \begin{pmatrix} \overrightarrow{\sigma}_{31} & \overrightarrow{\sigma}_{32} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \overrightarrow{\sigma}_{31}^{\,\prime} \\ \overrightarrow{\sigma}_{32}^{\,\prime} \end{pmatrix} \mathbf{\Sigma}_{33}^{-1} \begin{pmatrix} \overrightarrow{\sigma}_{31} & \overrightarrow{\sigma}_{32} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1} \\ \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1} \end{pmatrix} \begin{pmatrix} \overrightarrow{\sigma}_{31} & \overrightarrow{\sigma}_{32} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{31} & \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32} \\ \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{31} & \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{11} - \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{31} & \sigma_{12} - \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32} \\ \sigma_{21} - \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{31} & \sigma_{22} - \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32} \end{pmatrix} \\
&:= \begin{pmatrix} \sigma_{11|3\ldots p} & \sigma_{12|3\ldots p} \\ \sigma_{21|3\ldots p} & \sigma_{22|3\ldots p} \end{pmatrix} = \mathbf{\Sigma}_{12|3\ldots p}.
\end{aligned}
$$

Like the ordinary population correlation (e.g. Fujikoshi et al., 2010) between two variables X_1 and X_2 that is defined by

$$r(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}},$$

the correlation between the residuals $X_1^\star$ and $X_2^\star$ or rather partial correlation between X_1 and X_2 given $\overrightarrow{X}_3$ is

$$r(X_1^\star, X_2^\star) = \frac{\mathrm{Cov}(X_1^\star, X_2^\star)}{\sqrt{\mathrm{Var}(X_1^\star)\mathrm{Var}(X_2^\star)}} = \frac{\sigma_{12} - \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32}}{\sqrt{\left(\sigma_{11} - \overrightarrow{\sigma}_{31}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{31}\right)\left(\sigma_{22} - \overrightarrow{\sigma}_{32}^{\,\prime}\, \mathbf{\Sigma}_{33}^{-1}\, \overrightarrow{\sigma}_{32}\right)}} := \rho_{12|3\ldots p},$$

where the variances and covariance can be instantaneous taken from the variance/covariance matrix of $\overrightarrow{X}^\star$. At least we want to transform the partial correlation in a consistent form.

We define

$$\mathbf{\Sigma}^{-1}_{\cdot\cdot|1\ldots p\setminus\{\cdot\cdot\}} := \Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \ldots & \omega_{1p} \\ \omega_{21} & \omega_{22} & \ldots & \omega_{2p} \\ \vdots & & \ddots & \vdots \\ \omega_{p1} & \omega_{p2} & \ldots & \omega_{pp} \end{pmatrix},$$

also named *concentration matrix*. From the Inverse Variance Lemma (cf. Whittaker, 1990) it follows

$$\mathbf{\Sigma}^{-1}_{12|3\ldots p} = \left(\mathbf{\Sigma}^{-1}\right)_{[1:2,1:2]}$$

and hence,

$$\begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11|3\ldots p} & \sigma_{12|3\ldots p} \\ \sigma_{21|3\ldots p} & \sigma_{22|3\ldots p} \end{pmatrix}^{-1}$$

or turning it around leads to

$$\begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_{11|3\ldots p} & \sigma_{12|3\ldots p} \\ \sigma_{21|3\ldots p} & \sigma_{22|3\ldots p} \end{pmatrix}.$$

If the whole expression is inverted we yield

$$\begin{pmatrix} 1 & \star \\ -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \star \\ \frac{\sigma_{12} - \vec{\sigma}_{31}^{\,\prime}\mathbf{\Sigma}_{33}^{-1}\vec{\sigma}_{32}}{\sqrt{\left(\sigma_{11} - \vec{\sigma}_{31}^{\,\prime}\mathbf{\Sigma}_{33}^{-1}\vec{\sigma}_{31}\right)\left(\sigma_{22} - \vec{\sigma}_{32}^{\,\prime}\mathbf{\Sigma}_{33}^{-1}\vec{\sigma}_{32}\right)}} := \rho_{12|3\ldots p} & 1 \end{pmatrix},$$

where $\star$ is a place marker for an irrelevant entry of the matrix due to our aim. Since the expression in the right matrix is the partial correlation between X_1 and X_2 given $\vec{X}_3$, $\rho_{12|3\ldots p}$ can be written as

$$\rho_{12|3\ldots p} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}}.$$

An estimation of the partial correlation is given by

$$\hat{\rho}_{12|3\ldots p} = -\frac{\hat{\omega}_{12}}{\sqrt{\hat{\omega}_{11}\hat{\omega}_{22}}}.$$

As we discovered, the elements of the inverse covariace matrix are related to the partial correlation. Hence, a reliable estimation of the covariance matrix is required.

4.1.3 Shrinkage estimation of the covariance matrix

Especially in situations of "small n, large p", when much more variables than observations shall be considered, neither the maximum likelihood estimate $\boldsymbol{S}^{ML}$ nor the unbiased empirical covariance matrix $\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}^{ML}$ with

$$s_{il} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jl} - \bar{x}_l\right),$$

are good approximations of the true covariance matrix $\boldsymbol{\Sigma}$, not even if the number of variables and observations are approximately the same (Schäfer and Strimmer, 2005a). A property of the true covariance matrix is positive definiteness, if we assume that all considered random variables have non-zero variances. Yet, neither the unbiased nor the maximum likelihood estimator satisfy this requirement. Besides, it is desirable that a good covariance estimator is well-conditioned, i.e. the ratio of its minimum and maximum singular value is quite large, it has full rank and therefore, it can be easily inverted. This characteristic can be found in general neither in the maximum likelihood nor in the unbiased estimator.
A general approach to improve a covariance estimator is to reduce its variance. The mean squared error (MSE) of the sample covariance can be decomposed to variance and bias, i.e.

$$\mathrm{MSE}(\boldsymbol{S}) = \mathrm{Bias}(\boldsymbol{S})^2 + \mathrm{Var}(\boldsymbol{S}).$$

Since $\boldsymbol{S}$ is unbiased by construction, the overall accuracy of the unbiased estimator can only increase by the reduction of variance. Various approaches have been proposed for this issue and all procedures have serious disadvantages, e.g. to reduce the variance by bootstrap aggregation of the empirical covariance matrix (cf. Schäfer and Strimmer, 2005b) that becomes computationally highly expensive with increasing numbers of variables. A computationally inexpensive and simultaneously well performing "shrinkage" or rather "biased estimation" approach is described by Schäfer and Strimmer (2005a). It is based on the theorem of Ledoit and Wolf (2003), which is now briefly introduced.

Let $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_p)$ be the parameters of a high-dimensional unrestricted model of interest, and let $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_p)$ denote the matching parameters of a restricted lower

dimensional submodel, for instance parameters might be all equal, i.e. $\theta_1 = \ldots = \theta_p$. The estimates of $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are denoted by $\boldsymbol{U}$ and $\boldsymbol{Y}$, respectively. $\boldsymbol{Y}$ is also called shrinkage target. Due to the large number of parameters it is obvious that the unbiased estimate $\boldsymbol{U}$ will have a comparatively high variance, but $\boldsymbol{\Phi}$ might have a considerable high bias. In a linear shrinkage approach both estimates are combined to a new estimator

$$\boldsymbol{U}^* = \lambda \boldsymbol{Y} + (1 - \lambda)\, \boldsymbol{U},$$

where $\lambda \in [0, 1]$ is the shrinkage intensity that has to be selected. Apparently, for $\lambda = 0$ the regularized estimate $\boldsymbol{U}^*$ is equal to the unbiased estimator, while for $\lambda = 1$ the shrinkage target $\boldsymbol{Y}$ is recovered. This combined estimator can outperform the unbiased as well as the constrained estimator in terms of accuracy and efficiency.
Besides the choice of the shrinkage target the selection of the optimal shrinkage intensity remains. For the selection of the latter various procedures have been proposed. It is possible to fix the shrinkage intensity to a given value or a function that depends on the sample size. A computationally expensive approach for an optimal λ is cross-validation (e.g. see Friedman, 1989). In an empirical Bayes context, $E(\boldsymbol{Y})$ is interpreted as prior mean and λ as a hyper-parameter that has to be estimated from the data by optimizing the marginal likelihood (cf. Daniels and Kass, 2001). In this thesis we use a procedure where λ is choosen in a data-driven way by minimizing a risk function, here the mean squared error (MSE), which can be transformed to

$$\begin{aligned}
R(\lambda) &= E(L(\lambda)) \\
&= E\left(\sum_{i=1}^{p} (u_i^* - \phi_i)^2\right) \\
&= \sum_{i=1}^{p} \mathrm{Var}(u_i^*) + [E(u_i^*) - \phi_i]^2 \\
&= \sum_{i=1}^{p} \mathrm{Var}\left(\lambda y_i + (1-\lambda)\, u_i\right) + \left[E\left(\lambda y_i + (1-\lambda)\, u_i\right) - \phi_i\right]^2 \\
&= \sum_{i=1}^{p} \lambda^2 \mathrm{Var}\left(y_i\right) + (1-\lambda)^2\, \mathrm{Var}\left(u_i\right) \\
&\quad + 2\lambda\,(1-\lambda)\, \mathrm{Cov}\left(u_i, y_i\right) + \left[\lambda E\left(y_i - u_i\right) + \mathrm{Bias}\left(u_i\right)\right]^2 .
\end{aligned}$$

By minimizing this expression with respect to λ, we yield

$$\lambda^* = \frac{\sum_{i=1}^{p} \left[\mathrm{Var}\left(u_i\right) - \mathrm{Cov}\left(y_i, u_i\right) - \mathrm{Bias}\left(u_i\right) E\left(y_i - u_i\right)\right]}{\sum_{i=1}^{p} E\left[\left(y_i - u_i\right)^2\right]}.$$

From this expression we can derive that λ^* becomes small if the variance of $\boldsymbol{U}$ decreases. We see, the shrinkage target $\boldsymbol{Y}$ looses its influence when sample size increases. Furthermore, the correlation between $\boldsymbol{U}$ and $\boldsymbol{Y}$ influences the shrinkage intensity. If the two are positively correlated λ^* decreases as well as if the mean squared difference between $\boldsymbol{U}$ and $\boldsymbol{Y}$ increases which protects the regularized estimator $\boldsymbol{U}^*$ against erroneously chosen shrinkage targets. Moreover, if $\boldsymbol{U}$ is biased towards the shrinkage target $\boldsymbol{Y}$, λ^* will decrease. However, we assumed that $\boldsymbol{U}$ is unbiased. Then the equation above reduces to

$$\lambda^* = \frac{\sum_{i=1}^{p} \mathrm{Var}\left(u_i\right) - \mathrm{Cov}\left(y_i, u_i\right)}{\sum_{i=1}^{p} E\left[\left(y_i - u_i\right)^2\right]}.$$

For the estimation of λ^* Schäfer and Strimmer (2005a) propose to replace all expectations, variances and covariances by their unbiased sample counterparts, which leads to

$$\hat{\lambda}^* = \frac{\sum_{i=1}^{p} \widehat{\mathrm{Var}}\left(u_i\right) - \widehat{\mathrm{Cov}}\left(y_i, u_i\right) - \widehat{\mathrm{Bias}}\left(u_i\right)\left(y_i - u_i\right)}{\sum_{i=1}^{p} \left(y_i - u_i\right)^2}.$$

To keep λ^* in $[0, 1]$, it must be truncated afterwards to $\hat{\lambda}^{**} = \max\left\{0, \min\left\{1, \hat{\lambda}^*\right\}\right\}$. Transferring the lemma of Ledoit and Wolf (2003) to the covariance estimation issue we yield in matrix setting

$$\begin{aligned} L\left(\lambda\right) &= \left\|\boldsymbol{S}^* - \boldsymbol{\Sigma}\right\|_F^2 \\ &= \left\|\lambda\boldsymbol{Y} + \left(1 - \lambda\right)\boldsymbol{S} - \boldsymbol{\Sigma}\right\|_F^2 \\ &= \sum_{i=1}^{p}\sum_{l=1}^{p} \left(\lambda y_{il} + \left(1 - \lambda\right) s_{il} - \sigma_{il}\right)^2, \end{aligned}$$

where $\left\|\cdot\right\|_F^2$ denotes the Frobenius norm which is the eqivalent to the squared error loss function in matrix setting.

Finally the choice of the covariance shrinkage target is still pending. In Schäfer and Strimmer (2005a) several suitable shrinkage targets are presented. In this thesis we use

the "diagonal, unequal variance" target, where

$$y_{il} = \begin{cases} s_{ii} & , \ i = l \\ 0 & , \ i \neq l, \end{cases}$$

which shrinks only the off-diagonal elements of $\boldsymbol{S}$, and does not shrink the variances. Thus, $\hat{\lambda}^*$ reduces to

$$\hat{\lambda}^* = \frac{\sum_{i \neq l} \widehat{\mathrm{Var}}\left(s_{il}\right)}{\sum_{i \neq l} s_{il}^2}$$

since s is unbiased, the covariances $\widehat{\mathrm{Cov}}\left(y_{il}, s_{il}\right) = 0, \forall i \neq l$ and in the denominator $y_{il} = 0, \forall i \neq l$.

Thereby, it is reasonable to parameterize the covariance matrix in terms of variances and correlations, instead of variances and covariances, i.e. $s_{il}^* = \rho_{il}^* \cdot \sqrt{s_{ii} \cdot s_{ll}}, \ i, l = 1, \ldots, p$, because this formulations has two advantages. On the one hand, the (partial) correlations derived from the resulting, regularized covariance matrix $\boldsymbol{S}^*$ are independent of scale and location transformations of the data matrix. And on the other hand the off-diagonal elements determining the shrinkage intensity are on the same scale. The corresponding $\hat{\lambda}^*$ is simplified to

$$\hat{\lambda}^* = \frac{\sum_{i \neq l} \widehat{\mathrm{Var}}\left(\rho_{il}\right)}{\sum_{i \neq l} \rho_{il}^2}.$$

The resulting shrinkage estimator for $\boldsymbol{\Sigma}$

$$\boldsymbol{U}^* = \hat{\lambda}^{**} \boldsymbol{Y} + \left(1 - \hat{\lambda}^{**}\right) \boldsymbol{S}$$

will have the desired properties. Since a convex combination of a positive definite matrix ($\boldsymbol{Y}$) and a positive semi-definite matrix ($\boldsymbol{S}$) leads to a positive definite matrix, $\boldsymbol{U}^*$ will be positive definite, too, and can be inverted.

The obtained regularized covariance estimator may be used to calculate partial correlations as introduced in Section 4.1.2. These form the basis for the Graphical Gaussian models that are also called as covariance selection models.

4.1.4 The local False Discovery Rate

The next critical part of inferring Graphical Gaussian Models is model selection, i.e. to determine whether an edge is absent (null edge) or present (non-null edge). In gene association networks we expect most of the edges, and therefore the partial correlations, to vanish (Schäfer and Strimmer, 2005b). Therefore, we assume the distribution of observed partial correlations $\hat{\rho}$ across edges is given as a mixture

$$f(\hat{\rho}) = \eta_0 \cdot f_0(\hat{\rho}, \kappa) + (1 - \eta_0) \cdot f_A(\hat{\rho}),$$

of the null distribution f_0 and the distribution of partial correlations corresponding to the actually existing edges f_A. η_0 is the (unknown) proportion of null, i.e. non-existing edges. The naturally longer tailed density f_A is here assumed to be a uniform distribution from -1 to 1. The proportion of null-edges η_0 can be determined from the data, for algorithms see Efron (2004) and Storey (2002), respectively. In this thesis, proportion of $\eta_0 = 0.95$ proposed by Schäfer and Strimmer (2005a) is assumed. The density of the absent egdes can be easily computed in a closed form by

$$\begin{aligned} f_0(\hat{\rho}, \kappa) &= (1 - \hat{\rho}^2)^{\frac{\kappa-3}{2}} \frac{\Gamma\left(\frac{\kappa}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{\kappa-1}{2}\right)} \\ &= |\hat{\rho}| \, Be\left(\hat{\rho}^2 \,;\, \frac{1}{2}, \frac{\kappa-1}{2}\right), \end{aligned}$$

that is given in Hotelling (1953), where $Be(\kappa \,;\, a, b)$ denotes the $\beta-$distribution and κ is the degree of freedom. If we consider a large sample setting with $n > p$, the degree of freedom is $\kappa = n - p + 1$. Thus, the number of observations n must be larger than the number of variables p as we can see from the formula. If $n < p$ the distribution has the same form as mentioned above but the degree of freedom is not a simple function of n and p and has to be estimated from the data (Schäfer and Strimmer, 2005a).
The posterior probability of a null edge given the observed partial correlation may be written as

$$P(\text{null edge}|\hat{\rho}) = \frac{P(\text{null edge}) \cdot P(\hat{\rho}|\text{null edge})}{P(\hat{\rho})} = \frac{\eta_0 \cdot f_0(\hat{\rho}, \hat{\kappa})}{f(\hat{\rho})},$$

which is defined as the local False Discover Rate (lFDR)

$$lFDR\left(\hat{\rho}\right) := \frac{\hat{\eta}_0 \cdot f_0\left(\hat{\rho}, \hat{\kappa}\right)}{f\left(\hat{\rho}\right)}$$

given the observed partial correlation for a specific edge (Schäfer and Strimmer, 2005a). Following Efron (2005) we assume an edge to be significant, i.e. present if its local FDR is smaller than 0.2.

4.1.5 Measures for the comparison of two undirected graphs

Our aim is to compare two genetic networks that are obtained from two conditions of two groups of patients. Therefore, both networks are estimated separately and afterwards the difference or similarity of the networks is determined by a suitable measure. In the following a collection of such measures is introduced.
Let $\vec{X} = (X_1, \ldots, X_p)^{'}$ be a vector of random variables with population correlation matrix $\hat{\boldsymbol{r}} = (r_{kj})_{k,j\in\{1,\ldots,p\}}$ and let $\hat{\boldsymbol{\rho}} = (\rho_{kj|\{1,\ldots,p\}\setminus\{k,j\}})_{k,j\in\{1,\ldots,p\}}$ be the matrix of estimated partial correlations. Denote the enties of the upper triangular matrix of estimated ordinary correlations and partial correlations by $\vec{r}_c = (\hat{r}_{ci})_{i=1,\ldots,E}$ and $\vec{\rho}_c = (\hat{\rho}_{ci})_{i=1,\ldots,E}$, respectively, where $c = 1, 2$ are the two conditions or groups and $E = \frac{p(p-1)}{2}$ is the number of possible egdes.
In the folloowing we introduce 14 measures for the comparison of two networks. An overview of these measures is given in Table 12.

Maximum absolute distance of partial correlations (MaxDApC)
For all edges the absolute deviations of the partial correlations between two networks is calculated. The MaxDApC quantifies the difference of two networks by considering only the largest distance of all estimated partial correlations between the two groups. A high value of the statistic

$$T_1\left(\vec{\rho}_1, \vec{\rho}_2\right) = \max_{i\in\{1,\ldots,E\}} |\rho_{1i} - \rho_{2i}|$$

argues for at least one difference between two networks under the regarded conditions. However, this measure does not take the number of differences into account. A moderate change in partial correlations through several edges might remain undiscovered.

	abbreviation	description
T_1	$MaxDApC$	Maximum absolute distance of partial correlations
T_2	$MDApC$	Mean absolute distance of partial correlations
T_3	$MDQpC$	Mean quadratic distance of partial correlations
T_4	MDE	Mean difference of edges
T_5	χE	χ^2 statistic based on edges
T_6	$MDAR$	Mean absolute distance of ranks
T_7	$MDQR$	Mean quadratic distance of ranks
T_8	$MDARE$	Mean absolute distance of ranks of present edges
T_9	$MDQRE$	Mean quadratic distance of ranks of present edges
T_{10}	$CORpCE$	Pearson correlation of partial correlations corresponding to present edges
T_{11}	$RCORpCE$	Spearman correlation of partial correlations corresponding to present edges
T_{12}	$MaxDAC$	Maximum absolute distance of ordinary correlations
T_{13}	$MDAC$	Mean absolute distance of ordinary correlations
T_{14}	$MDQC$	Mean quadratic distance of ordinary correlations

Table 12: Overview of measures for the comparison of two networks.

Mean absolute distance of partial correlations (MDApC)

For the calculation of the MDApC (cf. Gill et al., 2010) again all absolute deviations of the partial correlations between two networks are calculated. Then the MDApC is defined as the mean of all distances divided by the number of possible edges, i.e.

$$T_2\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{E}\sum_{i=1}^{E} \left|\rho_{1i} - \rho_{2i}\right| .$$

Like for MaxDApC high values of MDApC suggest differences between the two graphs. One modified partial correlation leads to alterations of partial correlations of adjacent nodes, thus MDApC might detect smaller differences between the networks.

Mean quadratic distance of partial correlations (MDQpC)

The MDQpC is quite similar to the MDApC but it considers the quadratic distances instead of absolute deviations, i.e.

$$T_3\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{E}\sum_{i=1}^{E} \left(\rho_{1i} - \rho_{2i}\right)^2 .$$

Therefore, we expect MDQpC to respond more sensitive to larger differences of partial correlations.

Mean difference of edges (MDE)
For the MDE the local FDR must be calculated for every edge first to decide whether an edge is present or absent. The MDE counts the differences of existing and non-existing edges and divides this number by the number of possibly existing edges to take the size of the network into account, i.e.

$$T_4\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{E}\sum_{i=1}^{E} |I(\rho_{1i}) - I(\rho_{2i})|$$

where

$$I(\rho_{ci}) = \begin{cases} 1 & , lFDR(\rho_{ci}) \leq 0.2 \\ 0 & , lFDR(\rho_{ci}) > 0.2 \end{cases}$$

denotes the indicator function for significant edges. This measure is straight forward but it is heavily dependent of the proposed threshold of 0.2. Obviously, a large value indicates large differences between the groups.

χ^2 statistic based on edges (χE)
After determining which edges are present or absent in the two networks denote the numbers as follows in a contingency table:

		network 1		
number of edges		present	absent	$\sum$
network 2	present	e_{11}	e_{10}	$e_{1\cdot}$
	absent	e_{01}	e_{00}	$e_{0\cdot}$
	$\sum$	$e_{\cdot 1}$	$e_{\cdot 0}$	E

From this table we are able to calculate the χ^2 statistic with Yates' continuity correction

which is defined as χE, i.e.

$$\tilde{T}_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{E \cdot \left(|e_{11} \cdot e_{00} - e_{10} \cdot e_{01}| - \frac{E}{2}\right)^2}{e_{1\cdot} \cdot e_{\cdot 1} \cdot e_{0\cdot} \cdot e_{\cdot 0}}$$

and use it as a measure of indepence for the networks of two groups. For the application of a χ^2-test requires the assumption of a discrete probability of observed binomial frequencies can be approximated by the continuous χ^2 distribution which introduces some bias (Yates, 1934). This error should be corrected by the modified statistic as implemented as default in R (R Core Team, 2013) which has advantages especially in case of small expected cell frequencies. A small value of χE argues for independence and thus, for differences between the groups. Hence, we set $T_5 := -\tilde{T}_5$ to unify the interpretation of the measures. However, the χE statistic only measures the deviance between observed and expected number of edges. Hence, it does not detect changes from positive to negative partial correlations, or vice versa, as long as the edge is present according to the local FDR.

Mean absolute distance of ranks (MDAR)
To receive a more robust measure, we order the absolute partial correlations independently for both groups and assign ranks to them. The MDAR is closely related to the MDApC, only the partial correlations are replaced by the corresponding ranks, i.e,

$$T_6\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{E}\sum_{i=1}^{E} \left|rk(\rho_{1i}) - rk(\rho_{2i})\right|.$$

where rk denotes ranks of the observations.

Mean quadratic distance of ranks (MDQR)
We define the MDQR analogue to the MDAR using quadratic distances instead of absolute deviations:

$$T_7\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{E}\sum_{i=1}^{E} \left(rk\left(\rho_{1i}\right) - rk\left(\rho_{2i}\right)\right)^2.$$

We might expect an advantage in recognizing larger differences compared to the MDAR. The interpretation of both statistics, the MDAR and MDQR, remains the same. High

values of them indicate large differences between the networks.

Mean absolute distance of ranks of present edges (MDARE)
The MDARE is highly related to the MDAR. For the MDARE we restrict to edges that are present in one or both networks, i.e.

$$T_8\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{\mid sig\mid} \sum_{i \in sig} \left| rk\left(\rho_{1i}\right) - rk\left(\rho_{2i}\right)\right|,$$

where *sig* denotes the set of significant, i.e. present edges in at least one graph. The idea is to eliminate irrelevant information that might cause noise.

Mean quadratic distance of ranks of present edges (MDQRE)
For the MDQRE we take the ranks calculated for MDARE and average the quadratic differences through all edges from the set significant edges

$$T_9\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{1}{\mid sig\mid} \sum_{i \in sig} \left(rk\left(\rho_{1i}\right) - rk\left(\rho_{2i}\right)\right)^2,$$

where *sig* denotes the set of significant, i.e. present edges in at least one graph as before.

Pearson correlation of partial correlations corresponding to present edges (CORpCE)
The CORpCE is defined as the ordinary Pearson correlation of the partial correlations restricted to the set of edges that are significant in at least one graph, i.e.

$$\tilde{T}_{10}\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{\sum_{i \in sig}\left(\rho_{1i} - \overline{\rho}_1\right) \cdot \left(\rho_{2i} - \overline{\rho}_2\right)}{\sqrt{\sum_{i \in sig}\left(\rho_{1i} - \overline{\rho}_1\right)^2 \cdot \sum_{i \in sig}\left(\rho_{2i} - \overline{\rho}_2\right)^2}}.$$

In contrast to most measures introduced above a high value of CORpCE argues for a strong similarity of the network under both conditions. Again, to unify the interpretation of the statistics, we set $T_{10} := -\tilde{T}_{10}$.

Spearman correlation of partial correlations corresponding to present edges (RCORpCE)
For a more robust measure we use the Spearman correlation coefficient on the significant

(i.e. present in at least one graph) edges and exclude all non-significant edges as before to avoid noise caused by non-differential edges. Thus, we can write RCORpCE as

$$\tilde{T}_{11}\left(\vec{\rho}_1, \vec{\rho}_2\right) = \frac{\sum_{i\in sig}\left(rk(\rho_{1i}) - \overline{rk(\rho_1)}\right)\cdot\left(rk(\rho_{2i}) - \overline{rk(\rho_2)}\right)}{\sqrt{\sum_{i\in sig}\left(rk(\rho_{1i}) - \overline{rk(\rho_1)}\right)^2\cdot\sum_{i\in sig}\left(rk(\rho_{2i}) - \overline{rk(\rho_2)}\right)^2}}.$$

Like for χE and CORpCE define $T_{11} := -\tilde{T}_{11}$ to ensure a consistent interpretation of all statistics.

All measures presented above (cf. Lohr et al., 2010) are based on partial correlations since Graphical Gaussian models outperform relevance networks, and hence, partial correlations have advantages compared to ordinary correlations because they able to recognize indirect interactions (Schäfer and Strimmer, 2005a). Nevertheless we will have a closer look on three measures based on ordinary correlations.

Maximum absolute distance of ordinary correlations (MaxDAC)

On the lines of the MaxDApC the maximum of all absolute deviations between the two groups is considered for MaxDAC but here the differences of the ordinary population correlationd are calculated, i.e.

$$T_{12}\left(\vec{r}_1, \vec{r}_2\right) = \max_{i\in\{1,\ldots,E\}}\left|r_{1i} - r_{2i}\right|.$$

Mean absolute distance of ordinary correlations (MDAC)

Also for the MDApC we define a population correlation based counterpart, named MDAC, that is defined by

$$T_{13}\left(\vec{r}_1, \vec{r}_2\right) = \frac{1}{E}\sum_{i=1}^{E}\left|r_{1i} - r_{2i}\right|.$$

Mean quadratic distance of ordinary correlations (MDQC)
Finally, analogously to the MDQpC we define

$$T_{14}\left(\vec{r}_1, \vec{r}_2\right) = \frac{1}{E}\sum_{i=1}^{E}\left(r_{1i} - r_{2i}\right)^2$$

as MDQC as a distance measure for the comparison of two networks.

4.1.6 Permutation tests for the quantification of differential networks

After the calculation of the above mentioned measures a criterion is required to assess whether the value of a specific statistic argues for differences between the two considered networks. An obvious strategy is to test the hypothesis of differential networks and compute p-values. Since none of the statistics follows an established distribution we need to simulate it via permutation test technique (cf. Lohr et al., 2010).

Let $\vec{X} = (X_1, \ldots, X_p)$ be a random vector and let $\vec{x}_1, \ldots, \vec{x}_n$ denote independent and identically distributed samples of this random vector $\vec{X}$. Further, we assume the Graphical Gaussian Model of group $c = 1$ (GGM_1) to be estimated by a sample $\vec{x}^1_1, \ldots, \vec{x}^1_{n_1}$ of $\vec{X}^1$ that is independent of that of the second group (GGM_2) estimated from $\vec{x}^2_1, \ldots, \vec{x}^2_{n_2}$ of $\vec{X}^2$, where $n_1 + n_2 = n$. Here, $\vec{X}^1$ and $\vec{X}^2$ may have different distributions. With permutation tests we intend to test the hypothesis H_0 : "GGM_1 and GGM_2 are identical" against H_1 : "GGM_1 and GGM_2 are not identical".

The permutation test procedure proceeds as follows: First, we estimate the networks GGM_1 from $\vec{x}^1_1, \ldots, \vec{x}^1_{n_1}$ and GGM_2 from $\vec{x}^2_1, \ldots, \vec{x}^2_{n_2}$, respectively, and compute the statistic of interest T_y, $y \in \{1, \ldots, 14\}$ to quantify the difference between the networks under the given conditions. For $h = 1, \ldots, H$ we pool the samples of both groups $\vec{x}^1_1, \ldots, \vec{x}^1_{n_2}, \vec{x}^2_1, \ldots, \vec{x}^2_{n_2}$ into a single sample and use a random permuation $\zeta_{[h]}$ to re-arrange the elements of the pooled sample $\zeta_{[h]}\left(\vec{x}_1, \ldots, \vec{x}_n\right)$. Afterwards the re-arranged sample is subdivided in two subsamples of the same sizes as the original samples $\vec{x}^{\zeta_{[h]}}_1, \ldots, \vec{x}^{\zeta_{[h]}}_{n_1}$

and $\vec{x}^{\zeta_{[h]}}_{n_1+1}, \ldots, \vec{x}^{\zeta_{[h]}}_{n}$. From these subsamples the statistic of interest $T_{y,h}$ is computed. After recording the H values from the permutations of the considered statistic compute the fraction q_y of permutation statistics $T_{y,h}$ that are greater or equal to the original statistic T_y:

$$q_y = \frac{1}{H+1} \sum_{h=1}^{H} I_y^{\dagger}(T_{y,h} \geq T_y),$$

where $I_y^{\dagger}(T_{y,h} \geq T_y) = 1$ if $T_{y,h} \geq T_y$, and zero otherwise. Since for all our statistics high values indicate differences between the networks of the two groups, we can interpret the fraction q_y as one-sided permutation test p-value.

4.2 Detection of differential genetic networks

One crucial point for differential network analysis is the selection of genes for which differences in the interaction structure can be assumed between two groups of individuals. In the following we present two different hypothesis generating concepts to detect differential genetic networks.

4.2.1 Predefined networks from literature

The first strategy is the examination of predefined gene groups that arises from biological knowledge. Large databases and bioinformatic initiatives like Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), or the Reactome database can be screened for differences in genetic networks.
The Reactome database is free, open-source, curated and peer reviewed, available on *http://www.reactome.org/*. It aims to provide "intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education" (Milacic et al., 2012, and Croft et al., 2008). The user is able to download pathways of the following categories for Homo sapiens: Apoptosis, Binding and Uptake of Ligands by Scavenger Receptors, Cell Cycle, Cell-Cell communication, Cellular responses to stress, Circadian Clock, Developmental Biology, Disease, DNA Repair, DNA Replication, Extracellular matrix organization, Gene Expression, Hemostasis, Immune System, Meiosis, Membrane Trafficking, Metabolism, Metabolism of proteins, Muscle contraction, Neuronal System, Reproduction, Signal Transduction, SUMOylation, and Transmembrane transport of small molecules. Further sub-categories in hierarchical order can be selected and downloaded in the required formats. Pathways are also provided for other organisms.
KEGG is a free database that provides tools for "understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information"(KEGG, *http://www.kegg.jp/kegg/*). It represents biological systems, and combines genomic and chemical information with systems information. KEGG consists of sixteen main databases, at which the KEGG Pathway database is the most relevant for us. It contains pathway maps in 7 main categories, namely "Metabolism", "Genetic Information Processing", "Environmental Information Processing", "Cellular Processes", "Organismal Systems", "Human Diseases",

and "Drug Development", for which hundreds of pathways for different organisms are available for download or online visualization.
The aim of the GO project is to unify the representation of genes and gene products across different organisms and databases (Ashburner et al., 2000). It is an international bioinformatic initiative to maintain and develop its controlled vocabulary and to annotate gene and gene product attributes and to provide tools for easy access to all aspects of the data provided by the project, available on *http://www.geneontology.org/*. GO consists of the onlogies "biological process", "molecular function", and "cellular component". The three ontologies contain genes and can be described by a directed acyclic graph, such that all downstream nodes are a subset of the upstream node above it. In this way the gene groups become more specific in descending hierarchical order.
Furthermore, some disease-specific databases, like the "Genes-to-Systems Breast Cancer Database" (G2SBC) can be found on the internet. The G2SBC database provides a collection of data about genes, transcripts and proteins which have been reported in literature to be altered in breast cancer cells and includes mathematical models on cancerogenesis, tumour growth and tumour response to treatments (Mosca et al., 2010). It provides breast cancer genes, common molecular alterations in breast cancer, common KEGG pathways and enriched GO terms if a pathway or GO group is assumed to be breast cancer related. However, one needs to have some prior knowledge which gene group might by interesting. On cellular systems level it is also possible to assess lists of genes that are related to phenotypes, e.g. "grade 1(2) vs. 3". Another interesting aspect of this database is the section "Mathematical models related to cancerogenesis, tumour growth and response to treatments", but this section is quite obsolete, since all models were published between 1995 and 2007.

Instead of genome-wide screening for gene groups of pathways that differ between two conditions or phenotypes, a literature search for gene groups previously identified as phenotype-related made sense. e.g. in squamous lung cancer Wang (2012) found the GO groups "GO:0005576" (Extracellular region), "GO:0050828" (Regulation of liquid surface tension), and other GO terms to be significantly metastasis-related. The term "MHC protein complex" (GO:0042611) was shown to be cancer related in different tumor tissues. Gene signatures known to be associated with some phenotype can also be a starting point, e.g. Liu et al. (2007) found a 186-gene-signature that

predicts the invasiveness of breast tumors, Invshina et al. (2006) present a 264-gene-signature for the prediction of the histological grade, and van't Veer et al. (2002) developed a 70-gene-signature, known as "MammaPrint" genes commercialised by Agendia (*http://www.agendia.com/pages/mammaprint/21.php*) to predict prognosis (Tian et al., 2010).

We take the MammaPrint genes as basis for further investigations of differences in interaction networks between breast cancer patients of the Mainz cohort (cf. Section 2.2.1). 86 probe sets that correspond to 52 genes are present on the Affymetrix HG-U133A array, i.e. 18 genes are not represented by any probe set. Since the histological grade is highly correlated with prognosis, we split the patients according to tumor grade. 151 patients with grade I or II are grouped to the first class and 49 patients with grade III to a second class.
In van't Veer et al. (2002) it was shown that the MammaPrint genes are pognostic, however, it cannot be supposed that these genes build a genetic network with changing interactions. Therefore, we performed a gene set enrichment analysis by testing the independence of the two events 1. gene i is in the list of (interesting) MammaPrint genes and 2. gene i is a member of GO term with Fisher's exact test (e.g. Lehmann and Romano, 2005) which is implemented in the R package topGO (Alexa et al., 2006). The p-value depicts the probability of observing at least the same amount of enrichment when interesting genes are randomly selected out of all genes. Hence, a small p-value gives strong evidence for an over-representation of MammaPrint genes in a specified GO term. Applying this test to all GO terms of all three ontologies that contain 10 to 100 probe sets, we yield 58 GO groups with a raw p-value < 0.01, i.e. MammaPrint genes are over-represented in 40 GO terms of the ontology biological process, in 11 terms of the molecular function ontology, and in 7 GO terms of the ontology cellular component.
To the 58 GO terms enriched with MammaPrint genes we apply permutation tests using

significant tests	0	1	2	3	4	5	6	7	8
GO terms	23	7	14	2	7	2	1	1	1

Table 13: Number of significant tests (referred to $\alpha = 0.05$) for 58 enriched GO terms for MammaPrint genes.

the statistics introduced in Section 4.1.5. Table 13 gives an overview of the frequencies

of significant permutation tests for the considered GO terms. 1000 permutations were conducted. The range goes from 0 to 8 significant tests per GO group. For 23 GO terms no test is significant on an α-level of 5%. Applying the permutation tests to the original MammaPrint genes significance can be observed for the tests using MDAR, MDAC, and MDQC. An overview of the most noticable GO groups with description of the term, number of probe sets and the associated ontology can be found in Table 14.

GO term	description	probe sets	ontology	sign. tests
GO:0032332	positive regulation of chondrocyte differentiation	22	BP	8
GO:0070628	proteasome binding	12	MF	7
GO:0031663	lipopolysaccharide-mediated signaling pathway	59	BP	6
GO:0031532	actin cytoskeleton reorganization	70	BP	5
GO:0008608	attachment of spindle microtubules to kinetochore	25	BP	5

Table 14: Overview of most noticeable of the enriched GO terms for MammaPrint.

test	sig. GO terms
MaxDApC	0
MDApC	8
MDQpC	9
MDE	3
χE	5
MDAR	8
MDQR	9
MDARE	3
MDQRE	3
CORpCE	6
RCORpCE	6
MaxDAC	4
MDAC	16
MDQC	16

Table 15: Number of significant GO groups (referred to $\alpha = 0.05$) of 58 enriched GO terms for MammaPrint genes for all considered tests.

In Table 5 in the appendix the results of the 14 tests for all 58 GO terms are shown. All p-values are unadjusted and therefore considered as descriptive measures. The tests using MDAC and MDQC, MDApC and MDQpC, MDAR and MDQR, MDARE and MDQRE, and CORpCE and RCORpCE as test statistics agree in significance on an

α-level of 5% for most GO terms, but this is due to their similar design. For 16 GO groups the tests using MDAC and MDQC can reject the null-hypotheses of no differences between the networks of patients with tumor grade I or II and grade III, which makes MDAC and MDQC the tests with most significant findings by far (cf. Table 15). In contrast, with the test using MaxDApC we yield no noticeable GO group. Of course, the properties of the proposed tests must be analysed, i.e. if the tests hold a given α-level, which is done in Section 4.3.

4.2.2 DiNGS - Gene selection for differential networks

In this section we present a novel approach for the detection of differential networks between two groups of patients. The idea is to search iteratively for features that maximize the difference of the resulting network under two conditions or between two groups. That is we perform a kind of forward selection adapted for differential networks, similar to model selection for regression models by AIC (Akaike, 1974) or BIC (Schwarz, 1978). Penalized regression, like ridge regression (Hoerl and Kennard, 1970) or Lasso regression (Tibshirani, 1996) have been applied for inferring Graphical Gaussian Models by Meinshausen and Bühlmann (2005) that could also be implicitly used for model selection. We need to go one step further, because our aim is to detect *differential* networks. For that purpose we introduce a heuristic **Di**fferential **N**etwork **G**ene **S**election (DiNGS) algorithm. The general DiNGS proceeding with a default setting that is subsequently applied to the Mainz breast cancer dataset is shown in Figure 11. Next each step is descripted in detail.

Input: Gene expresssion dataset, criterion to split the patients in two groups

Output: Differential probe set network

Algorithm:

Step 1. Preselection of probe sets

Reduce dataset by selecting e.g. 100 probe sets with highest variance

Step 2. Definition of a starting probe set or starting probe set-pair

For $k = 1, \ldots, R$*:*

- *Sample a fraction of the reduced number of preselected probe sets and calculate partial correlations for each group;*
- *determine the minimal partial correlation per edge for each group and compute the distance of partial correlations per edge;*
- *define probe sets with maximum distance of partial correlations as start pair*

Step 3. Addition of probe sets

Find probe set with largest distance of partial correlations between the two groups of patients out of the remaining probe sets and add this probe set to the previously selected probe sets

Step 4. Assessment of difference between the selected probe set networks

Calculate the partial correlations of the current probe set selection for both groups of patients and build the mean squared difference

Step 5. Criterion to stop the algorithm

If the mean squared difference is below a threshold e.g. $v_u = 0.1$ *stop and reject the candidate probe set, else go to Step 3.*

Figure 11: DiNGS algorithm with standard settings as subsequently used on Mainz breast cancer dataset.

1. Selecting a differential network out of 20 000 or even 50 000 probe sets is challenging

and the computation of (partial) correlation matrices will be computationally expensive or even impossible. Hence, a *preselection of probe sets* is required. This can be done in many ways. One straightforward idea is to take probe sets with the highest variance across all samples. Thus, it is guaranteed to avoid genes or rather probe sets that are not expressed at all. To ensure selecting probe sets whose expression differs between both groups on the basis of variance, we propose to choose the probe sets with smallest values of

$$V(\text{probe set}_i) = \frac{\text{Var}_1\left(\text{probe set}_i\right) + \text{Var}_2\left(\text{probe set}_i\right)}{\text{Var}\left(\text{probe set}_i\right)},$$

where $\text{Var}_k\left(\text{probe set}_i\right),\ k = 1, 2$ denotes the variance of probe set i in group k. Another concept would be to cluster the probe sets e.g. by k-means or PAM (Kaufman and Rousseeuw, 2005) with $(1 - \text{correlation})$ as distance matrix and take the cluster with the highest average correlation as preselected group. Of couse, the number of clusters must be selected carefully. The best way might be to define a preselected set due to biological prior knowledge, e.g. be a larger sized GO term or KEGG pathway, where differences in a well-defined path or subset are expected. Furthermore, the effects of different preselection methods are described and discussed in the bachelor thesis from Cyris (2011).

2. After a subset of probe sets is selected we need to define one *probe set or a pair or probe sets to start* with, i.e. for building the network around it. Again, the probe set with the highest variance could be taken for that purpose, or analogously to the preselection the probe set with smallest within variance in the groups compared to the overall variance $V(\text{probe set}_i)$ to ensure differences between the two groups. Certainly, the definition of a starting gene or rather probe set by biological prior knowledge is possible. In contrast to the three approaches mentioned above, the following two criteria will lead to a starting *pair* instead of a single probe set. We propose to start with the pair with the highest correlation or, trying to take the difference of the groups into account, the pair with the maximal difference of minimal partial correlations. For assessment of minimal partial correlations, we condition the correlation on a subset of the preselected probe sets, e.g. $R = 100$ times (cf. Step 2 in DiNGS algorithm in Figure 11). The minimal partial correlation per edge out of 100 repeats is recorded, because we assume that this describes the actual correlation after removing the effect of all other influences at best.

3. *Addition of probe sets*: For the actual forward selection another criterion is required. Probe set i could be added to the previously selected group if it has the highest correlation or partial correlation conditioned on the previously selected probe sets either across all samples or in a reference group, e.g. healthy people. Again, the measure $V(\text{probe set}_i)$ can be used to extend the gene set. Another option is to select the $(s_i + 1)$-th probe set for the network by maximizing the sum of distances of partial correlations conditioned on all s_i previously selected probe sets

$$\max_{s_{i+1}} \sum_{i=1}^{E} |\rho_{1i} - \rho_{2i}| ,$$

where $E = \frac{(s_{i+1}-1)s_{i+1}}{2}$ denotes the number of possible edges of the advanced network. This criterion guarantees to add the node that maximizes the overall difference. The maximization of the distance between partial correlations of the candidate node with an already included node could also be considered, thus difference of previously affiliated probe sets may decrease through the influence of the recent node.

4. *To assess the difference between the selected probe set network* the MaxDApC, the MDApC, but also one of the other statistics for the quantification of difference proposed in Section 4.1.5 can be deployed. Another option is to consider as before the differences of edges concerning only the recently affiliated probe set.

5. Finally, a *criterion to stop the algorithm* on the basis of a measure to assess the difference (chosen before in 4.) is required. We might stop the algorithm and take the current set of probe sets for further investigations if none of the remaining probe sets lead to a difference of (partial) correlations above a threshold υ_u, which must be specified. One could also stop if the used statistic, determined in step 4, drops below a cut point υ_l or the previously determined maximal number of probe sets is achieved. A more time-expensive method would be to permute the data and compare the original findings with the random results in terms of average of maximal partial correlations.

Of course, not all combinations of the mentioned approaches are suitable for the detection of differential genetic networks and other combinations may be suitable for different

aims. E.g. for detecting networks where partial correlations differ preferably between all edges we need to select measures that consider the average distance of partial correlations. If we are interested in finding a network that is maximal different in e.g. one path, but the interaction structure retained in both groups should also be considered, we should use an approach based on a maximal difference of partial correlations but add further nodes by smallest (partial) correlation. Examplified we show the results for one suitable combination and analyze the stability of findings (cf. the bachelor thesis of Windgassen, 2011).

As preselected group we take the 100 probe sets with highest variance in the breast cancer cohort from Mainz and split the patients in a group with metastasis within 5 years after surgery and a second group without metastasis which are followed-up for at least 5 years. The groups contain 28 and 136 patients, respectively. Although the number of 100 features is pretty small and it might not be necessary here, we sample $R = 500$ times 20 probe sets of the 100 and the minimal partial correlation is recorded per edge. Afterwards, the absolute differences per edge are calculated and the corresponding probe sets of the maximum distance are taken as start pair. Subsequently, the starting probe sets are deleted from the list of preselected features. The partial correlations for each of the remaining probe sets with the selected features are calculated separately for both groups of patients and we build the absolute differences of partial correlations between the groups. After adding the probe set that corresponds to the maximum distance across all partial correlations to the selected set of features, the mean squared difference of partial correlations between the groups is computed. This step is repeated until the mean squared difference of partial correlations drops below a predefined threshold of $\upsilon_l < 0.1$. Finding the probe set pair to start with is the only random process in this variant of the algorithm. To analyze the stability of this selection we draw stratified bootstrap samples. In general, bootstrapping is a statistical method to assess the precision of an estimate (Hastie et al., 2001). Here, we sample n times with replacement of n samples B times, where B is 200 in this thesis. Unstratified sampling might cause datasets containing only patients of one group. Hence, we sample n_1 and n_2 samples according to the sizes of the original groups for each bootstrap sample. Calculating the maximal distance of minimal partial correlations as described above, we record the frequencies of beeing part of the final set for each probe set. The most frequently selected probe sets (in at least 40% of the iterations) are listed in Table 16.

probe set	frequency	symbol	gene name
212094_at	0.7	PEG10	paternally expressed 10
212092_at	0.65	PEG10	paternally expressed 10
204351_at	0.475	S100P	S100 calcium binding protein P
205509_at	0.46	CPB1	carboxypeptidase B1 (tissue)
202018_s_at	0.45	LTF	lactotransferrin
214087_s_at	0.445	MYBPC1	myosin binding protein C, slow type
207430_s_at	0.435	MSMB	microseminoprotein, beta-
203535_at	0.43	S100A9	S100 calcium binding protein A9
209301_at	0.415	CA2	carbonic anhydrase II
206022_at	0.405	NDP	Norrie disease (pseudoglioma)
206457_s_at	0.395	DIO1	deiodinase, iodothyronine, type I
209278_s_at	0.385	TFPI2	tissue factor pathway inhibitor 2
205242_at	0.38	CXCL13	chemokine (C-X-C motif) ligand 13
218332_at	0.38	BEX1	brain expressed, X-linked 1
218002_s_at	0.37	CXCL14	chemokine (C-X-C motif) ligand 14
209612_s_at	0.36	ADH1B	alcohol dehydrogenase 1B (class I), beta polypeptide
203290_at	0.355	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
37892_at	0.355	COL11A1	collagen, type XI, alpha 1
205357_s_at	0.35	AGTR1	angiotensin II receptor, type 1
214079_at	0.345	DHRS2	dehydrogenase/reductase (SDR family) member 2
213492_at	0.34	COL2A1	collagen, type II, alpha 1
222379_at	0.34	KCNE4	potassium voltage-gated channel, Isk-related family, member 4
203355_s_at	0.33	PSD3	pleckstrin and Sec7 domain containing 3
205513_at	0.33	TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)
219768_at	0.33	VTCN1	V-set domain containing T cell activation inhibitor 1
204475_at	0.325	MMP1	matrix metallopeptidase 1 (interstitial collagenase)
205916_at	0.325	S100A7	S100 calcium binding protein A7
205239_at	0.31	AREG	amphiregulin
213664_at	0.305	SLC1A1	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1
213831_at	0.305	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1

Table 16: Most frequently selected probe sets for final set obtained by stratified bootstrapping for the starting pair.

We see, two probe sets, namely "212094_at" and "212092_at" are most frequenctly selected by far. It is also the pair that is chosen as starting pair in 56% of all bootstrap samples. That is conspicuous, because both probe sets represent the same gene "paternally expressed 10" (PEG10) which encodes the retrotransposon-derived protein. The PEG10 gene includes two overlapping reading frames of the same transcript encoding distinct isoforms (Lux et al., 2005) and it is known to be overexpressed e.g. in hepatocellular carcinomas (Tsuji et al., 2011) and gallbladder adenocarcinoma (Liu et al., 2011). The change of partial correlation between metastatic and non-metastatic patients might indicate a change of association between different isoforms resulting from alternatively

spliced transcript variants of PEG10.
To summaries it, our selection approach based on the maximal difference of partial correlations adding further nodes by minimal partial correlation has two advantages: First, it is applicable also for larger preselected gene groups and second, through randomness different potentially interesting pairs may be found. The algorithm should be conducted several times.

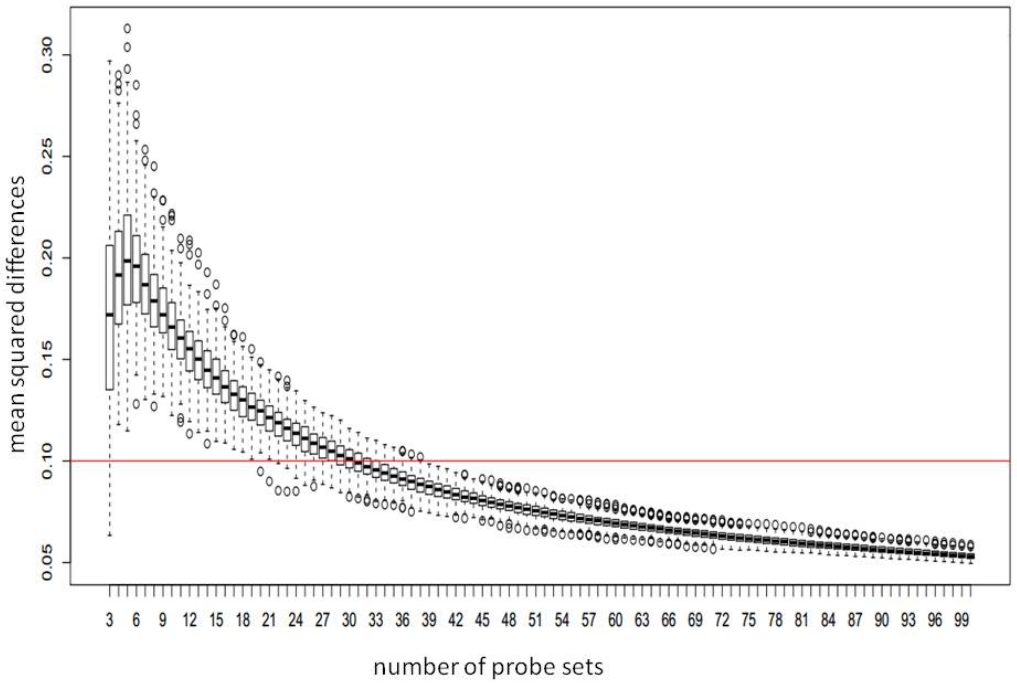

Figure 12: Results of the stratified bootstrap analysis for selection of differential genetic networks in the Mainz breast cancer cohort. Boxplots of the mean squared differences of partial correlations for each number of included probe sets are drawn. The red line indicates the boundary of 0.1 to stop at (reproduced from Windgassen, 2011).

In Figure 12 boxplots of the mean squared distance for all bootstrap samples ordered by the number of selected nodes are shown. The red line indicates the stop criterion. We see, mean and variance decrease with growing number of selected features. Most times the selected set contains between 19 to 38 features until it stops. It is conspicuous that the

algorithm stopped with merely the starting pair. Here, an improper pair of probe sets is selected to start with. But once another probe set is added to the starting pair the mean squared difference of partial correlations inceases above 0.1 for all bootstrap samples. Although the number of features varies from 19 to 38, the core probe sets remain the same. Thus, the variant of the DiNGS algoritm is just partial stable, it produces an appropriate number of features to apply further differential network analyses to.

4.3 A simulation study for the detection of differential neworks

To analyze the properties of the permutation tests with test statistics proposed in 4.1.5 we conduct an extensive simulation study. Therefore, we describe the design of simulated data and the settings first. Afterwards, we check whether the tests hold the α-level. Finally the power properties are explored.

4.3.1 Design of data

For the construction of data we take the well-studied RAF signalling pathway, also known as RAF-MEK-ERK pathway, (Sachs et al., 2005; Dougherty et al., 2005) that is often used as a gold standard network (e.g. Werhli et al., 2006). This signalling cascade describes the interaction of 11 phosphorylated proteins and phospholipids in human immune system cells. The central RAF protein is known to be involved in the regulation of cellular proliferation in immune cells through cell division cycle, apoptosis, cell differentiation, and cell migration. Dysfunctions in the regulation of the RAF pathway lead to uncontrolled growth and may cause proliferation in many cancers, e.g. melanomas and Hodgkin disease (cf. Zheng et al., 2003). Since several compounds are known that are able to inhibit various steps of the RAF signalling pathway, it is obvious to use this point of contacts as potential drug targets (e.g. Orton et al., 2005; Hilger et al., 2002; McCubrey et al., 2007). The ability to inhibit single compounds of the signalling cascade made it possible to infer the network structure via interventional data obtained using for example kinase-specific inhibitors (Sachs et al., 2005; Pearl et al., 2000).
The simplified network structure of the RAF signalling pathway is illustrated in Figure 13. The graph of this pathway consists of 11 nodes that are connected by 20 directed vertices. Albeit the proposed methods base on undirected network it is eligible and probably mandatory to construct directed data because our intention is to apply the proposed tests to real genetic networks and pathways that are necessarily directed.

Based on its network structure we are going to generate synthetic data to control the dependencies among the variables. To consider all dependencies correctly we need to assign a topological order. Therefor, we first introduce two definitions for directed graphs.

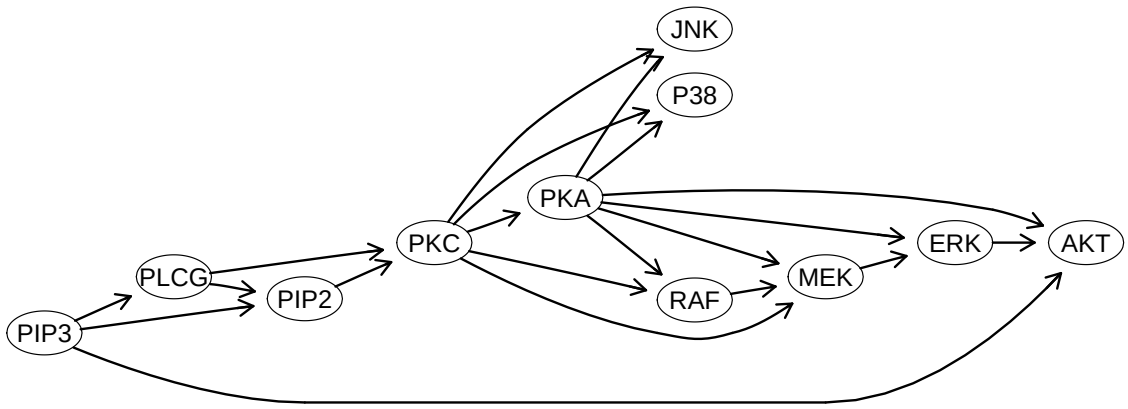

Figure 13: Model of the RAF signalling pathway.

For a directed graph we say node δ is a *parent* of γ if there is an edge from δ to γ (δ, γ) (cf. notation introduced in section 4.1.1) and $(\delta, \gamma) \neq (\gamma, \delta)$, $\delta, \gamma \in \mathcal{V}$, or the other way round, γ is said to be a *child* of δ. A *topological order* is an arrangement of nodes such that every node is ranked after all its parents (e.g. Lauritzen, 1996).
It is obvious that PIP3 is the only node in the graph that has no parents. Hence, PIP3 has to be top in a topological order. PLCG, PIP2, and AKT are the children of PIP3. The next node in a topological order has to be PLCG, because AKT has other two more parent nodes that are not yet ranked, and PIP2 is also a child of PLCG. If we continue, following this precept we might yield the topological order: PIP3−PLCG−PIP2−PKC−PKA−JNK−P38−RAF−MEK−ERK−AKT. This order is ambiguous, another possible order is PIP3 − PLCG − PIP2 − PKC − PKA − P38 − JNK − RAF − MEK − ERK − AKT, since JNK and P38 have the same parents.
According to the topological order we sample data from a linear Gaussian distribution by

$$X_i \sim N\left(\sum_{p_i} w_{ip_i}\tilde{x}_{p_i}, \sigma\right), i = 1, \ldots, 11$$

where the random variable X_i denotes the expression of node i with realizations $\vec{x}_i$, $N(\cdot)$ denotes a normal distribution, p_i are the parent nodes of node i, w_{ip_i} is the strength of interaction between node i and its parents nodes and $\tilde{x}_{p_i}$ are the standardized values as realisations of the random variable $\tilde{X}_{p_i}$ denoting the expression of the parent node p_i

with

$$\tilde{\vec{x}}_{p_i} = \frac{\vec{x}_{p_i} - \bar{\vec{x}}_{p_i}}{sd(\vec{x}_{p_i})}.$$

We standardize the values to avoid increasing variance along the topological order. σ is a noise term.

Sampling data from this model serveral parameters can be varied. As default, the interaction strength or rather coefficients w_{ip_i} are independently sampled from an uniform distribution over the interval $[0.5; 2]$ and provided with a randomly sampled sign following Werhli et al. (2006). We vary the noise in the data by setting the variance from small values $\sigma = 0.01$ to large values of 16. Since we intend to analyze tests for the difference of two networks we need to simulate networks for two groups and establish one or more differences between them. The sample sizes are considered as from small samples sizes with 20 observations per group to large balanced and unbalanced sample sizes with 200/200 and 300/100 observations per group, respectively, which are quite realistic sample sizes in real data. As differences the knockout of the central node PKC only or of the three nodes PIP2, PKC, and PKA, is considered and reported in this thesis. Other knockouts have been analyzed, but no major differences were found in principal. As knockout we mean the expression to disappear and therefor assume the corresponding underlying random variable or rather random variables to be normally distributed with mean zero which means no influence of other variables. Thus, all edges that pointed towards the knockout node will vanish.

Parameter	settings for simulation of data
variance of noise term	0.01; 0.1; 0.5; 1; 2; 4; 8; 16
sample sizes per group	100/100; 150/50; 180/20; 50/50; 20/20; 200/200; 300/100
knockout of node	PKC; (PIP2, PKC, PKA)
number of additional nodes	0; 5; 10; 20; 50

Table 17: Overview of settings for simulation of data based of the graph of the RAF signalling pathway.

In real data situations it is challanging to extract exactly these nodes belonging to the

network of interest as seen in Section 4.2. Hence, we generate a number of additional noise nodes whose expression assumed to be normally distributed with mean zero to the data. A complete overview of all parameter settings is given in Table 17.

All combinations of parameter settings are considered and for each combination 100 datasets are generated to analyze the properties of the permutation tests for the quantification of differences in two undirected networks. The results are described in the following section.

4.3.2 Properties of tests for the quantification of differential networks

Analyzing the α-level

A crucial point to know is whether a test holds the given significance level. Therefore, we simulate data without systematic differences, i.e. without any knockouts, between the two groups and apply the permutation tests proposed in Section 4.1.5 and 4.1.6. All combinations of settings listed in Table 17 in Section 4.3.1 except for the knockout of nodes are conducted. Considering all 100 datasets for every setting the proportion of rejected null-hypothesis for every significance level $\alpha_s \in [0\,;\,1]$ is determined. The decision of a test can be considered to be a random variable W with two feature characteristics. Let $W_1, \ldots, W_\Delta$ be sampling variables of W that are independent and identically distributed (i.i.d.) with

$$W_\zeta = \begin{cases} 1 & \text{, null-hypothesis is rejected} \\ 0 & \text{, null-hypothesis is not rejected,} \end{cases}$$

with $\zeta = 1, \ldots, \Delta$. Hence, W can be assumed to be Bernoulli distributed with success probability π, i.e. π is the probability of rejecting the null-hypothosis (if H_0 is true). It is known that the sum of Δ Bernoulli distibuted variables is binomial distibuted with parameters Δ and π, $\sum_{\zeta=1}^{\Delta} W_\zeta \sim Bin\,(\Delta,\, \pi)$. Therefore, we can test the hypothesis

$$H_0 : \pi \leq \alpha_s \qquad \text{against} \qquad H_1 : \pi > \alpha_s$$

using a Binomal test (e.g. Genschel and Becker, 2005). In the below-mentioned figures the upper boundary of the 95% confidence interval first given by Clopper and Pearson (1934) is plotted. By estimating and testing the proportion of rejected hypothesis we

test whether a permutation test in a specified scenario rejects too often under the null-hypothesis, i.e. it does not hold the given α-level.

Examplary, in Figure 14 the results of the scenarios with 20 samples per group and a high variance of 16 without any systematic differences between the groups are shown. The same plots for all other scenarios can be found in Figures 25 to 79 in the appendix. For every significance level $\alpha_s \in [0\,;\,1]$ on the x-axis, the corresponding proportion of rejected null-hypothesis, i.e. the estimated type I error, is drawn on the y-axis. Each plot corresponds to one of the permutation test statistics introduced in 4.1.5. The different colors indicate scenarios with different numbers of additional nodes, where light colors correspond to small numbers and dark colors to larger numbers of additional nodes. The red dashed line denotes the upper boundary of the 95% confidence interval for π described above.

For the permutation tests using the MaxDApC, MDApC, MDQpC, MDAR, MDQR, MaxDAC, MDAC, and MDQC as test statistics the proportion of rejected hypothesis under the null-hypothesis is less or equal to the upper boundary of the confidence interval for π in all scenarios. i.e. the mentioned tests do not reject the null-hypothesis too often, except for some random exceedings around significance levels around 0.5 mainly in permutation tests based on partial correlations. Noticeable are the runs of the curves of the MDE, χE, MDARE, and MDQRE as well as CORpCE and RCORpCE.

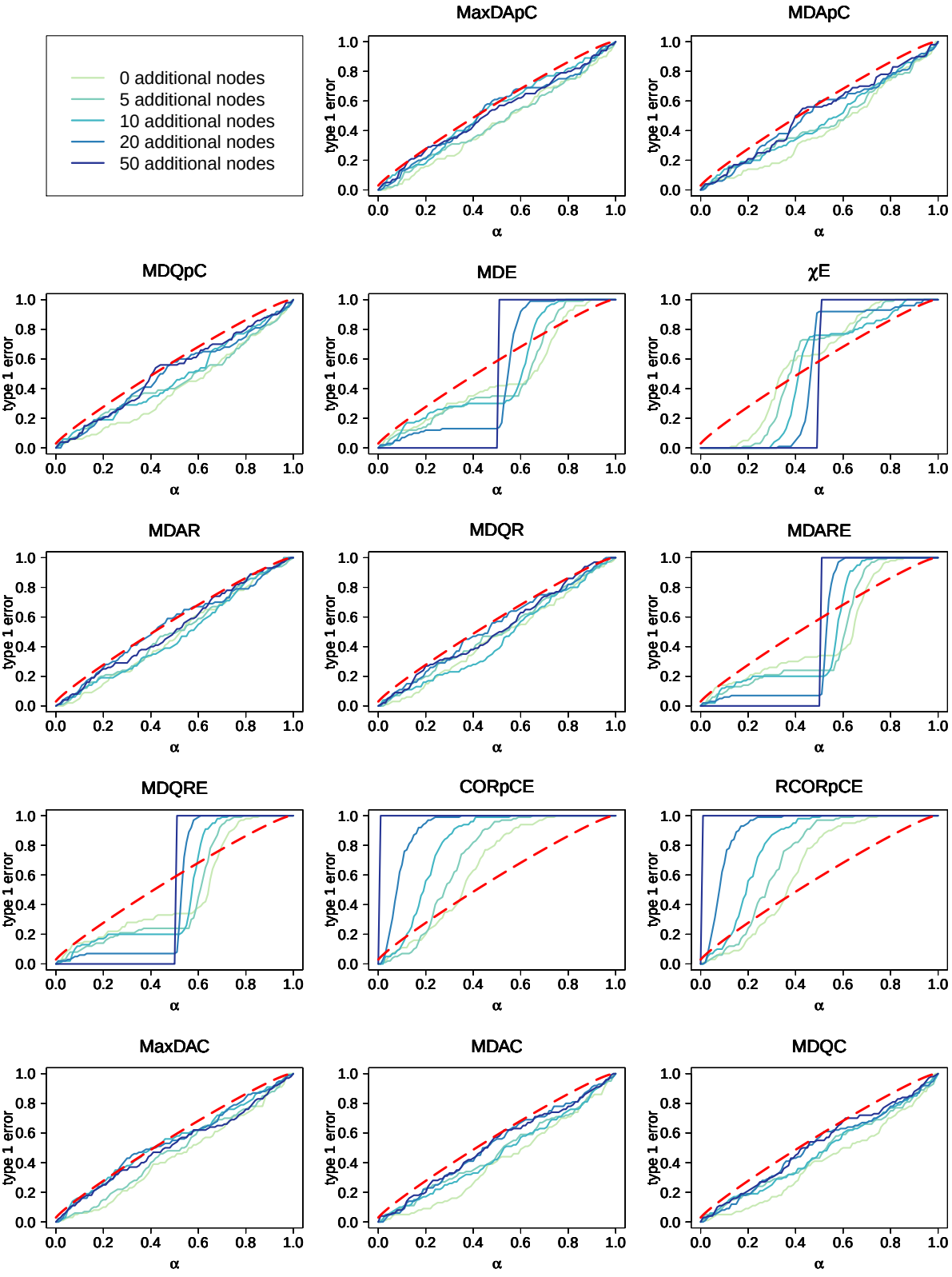

Figure 14: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples per group and noise 16.

Adding 50 nodes as described in Section 4.3.1 the proportion of rejected null-hypothesis is 0 from α_s between 0 and 0.5 for the permutation tests using MDE, χE, MDARE, and MDQRE. At an α-level close to 0.5 the curve sharply increases to 1, i.e. all hypotheses are rejected above this level. This means that all p-values are approximately 0.5. This is an artefact of the permutation tests with test statistics that use the local FDR and a fixed threshold to distinguish between significant and non-significant edges. Considering a local FDR of 0.2, no edges are present in the networks. With decreasing number of additional nodes, the rise of the curve is less abrupt because in some iterations we find present edges in one or both networks. Thus, the permutation tests with test statistics MDE, χE, MDARE, and MDQRE are improper, at least for small sample sizes and large number of nodes in the network.

Also the permutation tests using CORpCE and RCORpCE seem inappropriate for testing the difference of two networks with small samples sizes and a large number of vertices. As we see in Figure 14, the permutation tests reject the null-hypothesis much too often or, in case of 50 additional nodes, always, and therefore, they do not hold the significance level. The same results as for the CORpCE and RCORpCE permutation tests with 20 observations can be found for the scenario simulating 180 and 20 samples per group with a variance of 16 for the permutation test using the χE statistic (cf. Figure 48 in the appendix). Yet, in this scenario the χE permutation test does not hold the significance level for less or no additional nodes for smaller α's.

All permutation tests using statistics not mentioned above hold the significance level for every $\alpha_s \in [0\,;\,1]$. Especially the MDAC and MDQC permutation tests that depend on ordinary correlations instead of partial correlations do not exhaust the acceptable proportion of rejected null-hypotheses when the noise level and the number of nodes in the network decreases. This means, the permutation tests become more conservative with smaller noise level and a smaller number of nodes independent of the sample sizes.

General power properties

In this section we analyze the power properties of the proposed tests. We simulate 100

datasets for each setting described in Section 4.3.1 with knockout of node PKC or nodes PIP2, PKC, and PKA. Since the network is different by construction, we are under the alternative hypothesis. Hence, the proposed tests should reject the null-hypothesis at best 100 times. Such a test would have a power of 1 or 100%, but this seems unrealistic under all conditions of sample size and noise. To assess an estimate for the power of the proposed tests we count the fraction of rejected hypotheses of all tested hypotheses for each scenario. For rejecting a null-hypothesis a threshold of an α-level of 5% is determined. For the scenario where 100 samples per group are assumed and node PKC lost its parents the results can be found in Figure 15. The results of simulations assuming low variance of the noise term are drawn in light colors (yellow) and with increasing variance the power curves are marked in darker colors (blue). Since we merely simulated a discrete number of 0, 5, 10, 20, and 50 nodes, the filled dots represent the actual obtained power estimates. The curves are just rough interpolations for visualisation.

In general, the power decreases with growing number of additional nodes in the networks and higher variance of the noise term, while the power generally increases with the sample size (cf. Figure 15 and Figures 80 to 85 in the appendix). However there are some exceptions which are described in the following paragraphs.

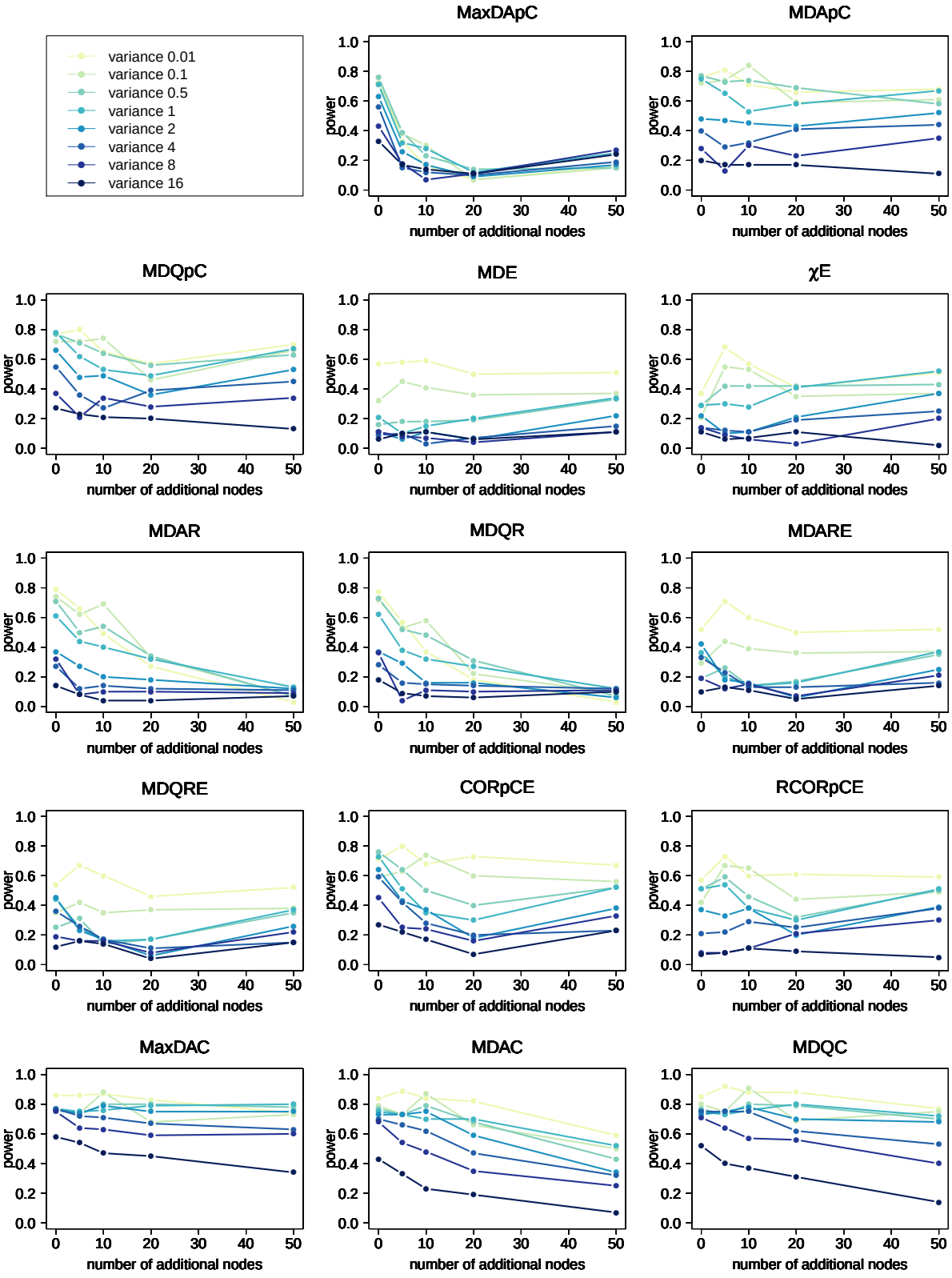

Figure 15: Proportion of rejected hypothesis for simulated setting of 100 samples per group and knockout of node PKC.

	variance of noise term 0.01 number of additional nodes					variance of noise term 16 number of additional nodes				
test statistic	0	5	10	20	50	0	5	10	20	50
MaxDApC	0.76	0.33	0.16	0.07	0.16	0.33	0.17	0.14	0.11	0.24
MDApC	0.76	0.81	0.71	0.66	0.68	0.20	0.17	0.17	0.17	0.11
MDQpC	0.77	0.80	0.65	0.57	0.70	0.27	0.23	0.21	0.20	0.13
MDE	0.57	0.58	0.59	0.50	0.51	0.06	0.10	0.11	0.06	0.11
χE	0.37	0.68	0.57	0.41	0.51	0.11	0.06	0.07	0.11	0.02
MDAR	0.79	0.66	0.49	0.27	0.03	0.14	0.08	0.04	0.04	0.07
MDQR	0.77	0.57	0.37	0.18	0.03	0.18	0.09	0.07	0.06	0.10
MDARE	0.52	0.71	0.60	0.50	0.52	0.10	0.13	0.11	0.05	0.14
MDQRE	0.54	0.67	0.60	0.46	0.52	0.12	0.16	0.14	0.04	0.15
CORpCE	0.71	0.80	0.68	0.73	0.67	0.27	0.22	0.17	0.07	0.23
RCORpCE	0.57	0.73	0.60	0.61	0.59	0.07	0.08	0.11	0.09	0.05
MaxDAC	0.86	0.86	0.87	0.83	0.74	0.58	0.54	0.47	0.45	0.34
MDAC	0.84	0.89	0.84	0.82	0.59	0.43	0.33	0.23	0.19	0.07
MDQC	0.85	0.92	0.88	0.88	0.77	0.52	0.40	0.37	0.31	0.14

Table 18: Proportion of rejected null-hypotheses of all considered tests if 100 samples per group and a noise term of 0.01 on the left side and a noise of 16 on the right side are assumed and a difference between the networks was created by eliminating the influence of the parents on node PKC. The rows corresponds to the different tests, the columns indicate the number of nodes added to the RAF-network structure as noise factor.

We see, assuming the parameters as specified above for Figure 15, that most permutation tests reach a power of approximately $70 - 80\%$ when no additional nodes and a small variance of the noise term are simulated. For small variances, the tests with MDE, χE, MDARE, MDQRE, and RCORpCE as test statistics have less power considering no additional nodes, but reach their maximal power when 5 or 10 nodes are added to the 11 nodes of the RAF-network. For larger variances of the noise term this effect cannot be observed. Table 18 shows, the estimated power for 100 samples per group and knockout of node PKC for the smallest and the highest considered variances of the noise term, 0.01 and 16, respectively.

The tests using MaxDAC, MDAC, and MDQC based on ordinary correlations have most power in both variance settings. But with growing number of additional nodes the MDAC loses more power than MaxDAC and MDQC. MDAR and MDQR perform quite well without additional nodes and low variance, but otherwise the power becomes very

small. The power of the test using CORpCE is the highest one of all considered tests that base on edges, i.e. where the local FDR is considered to decide whether an edge is present or absent.
In Figure 83 in the Appendix where the setting with only 20 samples per group is shown, we see that the power of the latter mentioned test increases dramatically with 50 additional nodes and higher variances. This is not a desirable property and arises from the fact that no edges are significant on a local FDR of 0.2 in both groups. The permutation test depending on MaxDApC rejectes about 75% of the null-hypotheses in the simplest setting with 100 samples per group, i.e. with variance of the noise term of 0.01 and no additional noise nodes. But the power decreases severely with increasing number of nodes in the network. MDApC and MDAQpC do not lose that much power with increasing number of additional nodes. But the power of the tests using their ordinary correlation based counterparts, namely MDAC and MDQC, is considerable higher, especially in settings with higher variances of the noise term.

In Figures 86 to 92 in the appendix the results of all settings where nodes PIP2, PKC, and PKA are modelled as normally distributed noise without influence of the corresponding parents are shown. Here, the differences are assumed to be larger than in the settings where only one node is knocked out, hence, we expect the tests to reject the null-hypotheses more often. Indeed, the power is generally higher.
Particularly, assuming 100 samples per group (cf. Figure 86 in the appendix), small variances of the noise term combined without any additional nodes the MaxDApC, MDApC, MDQpC, MaxDAC, MDAC, and MDQC tests have power of 100%. The MaxDAC tests even rejects all null-hypotheses up to a moderate noise level independently of the nodes added to the RAF-network structure. If no additional nodes are considered the tests using MaxDAC, MDAC, and MDQC have power above 90% for all variances of the noise term, while MaxDApC, MDApC, and MDQpC lose power with increasing noise. For MaxDAC, the power does not decrease below 60% in any setting where 100 samples per group and a knockout of nodes PIP2, PKC, and PKA are assumed. But without any additional nodes and high variance of the noise term the test using MDQC recognizes the difference between the networks a little more often. Moreover, the power is almost independent of the number of additional nodes for the tests using MaxDAC as test statistic if the amount of noise is not too high.

Decreasing power with unbalanced sample sizes

If we compare the results of the setting with 150 samples in the first and 50 samples in the second group where node PKC is knocked out (Figure 80 in the appedix) with the one where nodes PIP2, PKC, and PKA are modelled as normally distributed noise without influence of the corresponding parents (cf. Figure 87 in the appendix) the test using MaxDApC reached considerably more power than the MDApC and MDQpC tests in particular for larger numbers of additional nodes.
For both knockout settings the MDAC and MDQC tests have more power than their counterparts based on partial correlations, namely MDApC and MDQpC. The MaxDAC test has more power in almost all settings compared to the MaxDApC, except for the scenario when we simulate 50 additional nodes, high variance and knock out of nodes PIP2, PKC, and PKA. Here, the MaxDAC test has approximately 20% more power than the MaxDApC test.
The power curves of the settings with 180 samples in the first and 20 samples in the second group look pretty similar to the settings with 150 and 50 samples. In comparison to the settings with knockout of only node PKC the curves are shifted a little bit higher.

Comparing the power of the tests using MDApC and MDQpC of settings where unequal sample sizes are simulate with settings with smaller, but balanced group sizes, e.g. 150 and 50 samples with 50 samples in both groups, it becomes obvious that the tests have more power with smaller sample sizes. Examplified, we have a closer look on the permutation test using MDApC as test statistic on the setting with variance of the noise term of 1, and node PKC loses its parents (Figure 16). We compare the results of 300/100 samples to those of 100 samples per group. The power curves for 300/100 samples are contrasted with the ones for 100/100 in Figure 16. The 95% confidence interval for the proportion of rejected hypotheses

$$\left[\hat{\pi} - 1.96\sqrt{\frac{\hat{\pi}\cdot(1-\hat{\pi})}{\Delta-1}}\,;\,\hat{\pi} + 1.96\sqrt{\frac{\hat{\pi}\cdot(1-\hat{\pi})}{\Delta-1}}\right]$$

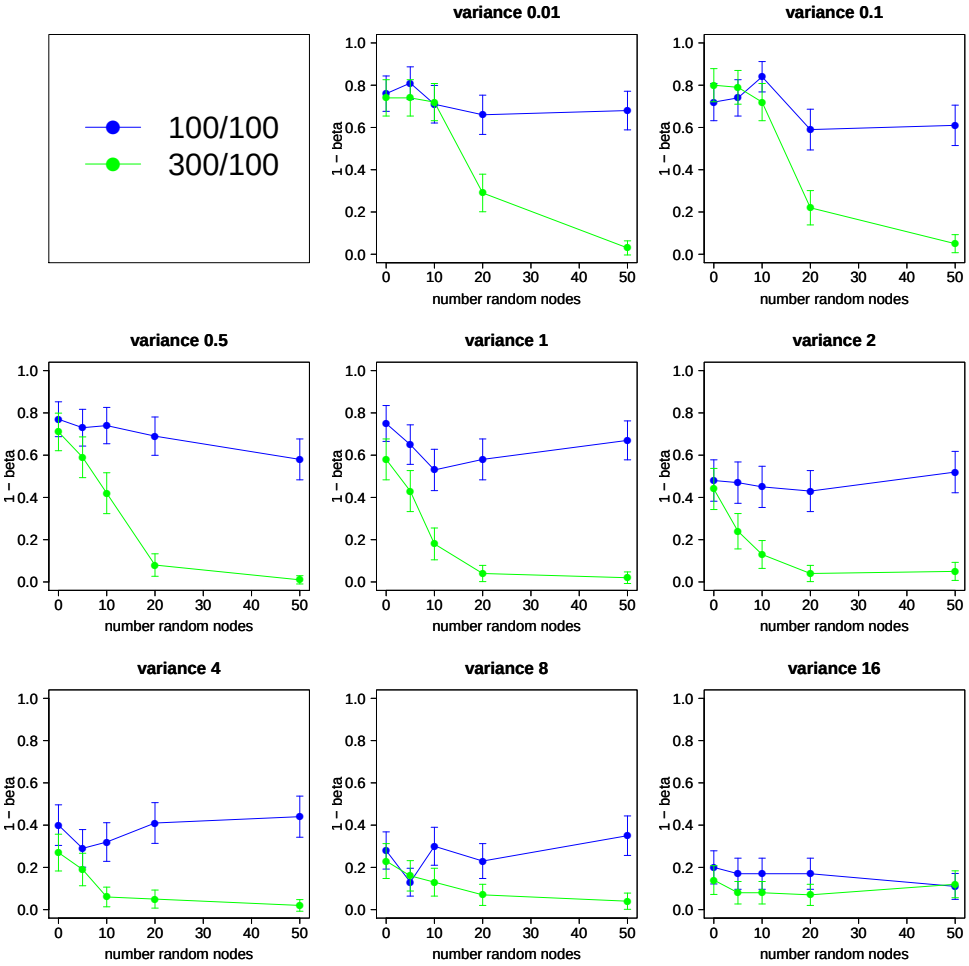

Figure 16: Comparison of power of the test using MDApC as test statistic between the scenario with 300 in the first and 100 in the second group and 100 samples in both groups (cf. caption Table 18).

is drawn for every setting in this figure, where Δ is the number of samples, here 100. We see, the amout of power difference grows with increasing numbers of nodes added to the original 11 nodes included in the RAF pathway. We have several speculations for this phenomenon.

It might be up to the avarage that is used for the MDApC statistic, because the test using MaxDApC is not affected. Another reason might be that partial correlations are used and we cannot see the effect on tests using ordinary correlations. And it might be due to the shrinkage of the covariance matrix that is used for the calculation of the partial correlations, because the effect does merely occur if unbalanced samples sizes are

considered.

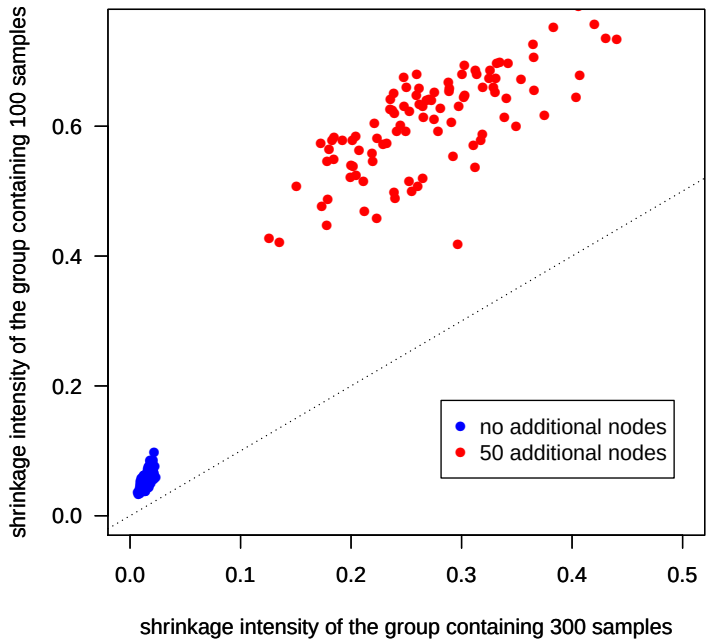

Figure 17: Comparison of shrinkage intensities in the scenario with 300 in the first and 100 samples in the second group without (blue) and 50 additional nodes (red), where variance of the noise term is 1, and node PKC is knocked out.

To understand the effect we have a look on the shrinkage intensities of a scenario with unequal sample sizes, first.

In Figure 17 the shrinkage intensities of the group containing 300 samples are drawn on the x-axis, and the shrinkage intensities of the group containing 100 samples on the y-axis. The blue colored dots correspond to the shrinkage intensities obtained from the dataset without any additional nodes, while red dots stand for intensities of the scenario where 50 nodes are added to the 11 nodes of the RAF network structure. We see, the shrinkage intensities of the group containing 100 samples are always greater than the one estimated from the data of 300 samples. With growing number of additional nodes the shrinkage intensities increase and the dots scatter wider. This is not only an effect of larger sample size in one group, it can also be observed with equal sample sizes (cf.

Figure 93 in the appendix). These results can be easily explained by transforming the formula for the estimation of the optimal shrinkage intensity:

$$\hat{\lambda}^{\star} = \frac{\sum_{i \neq j} \hat{Var}\left(r_{ij}\right)}{\sum_{i \neq j} r_{ij}^2} = \frac{\sum_{i \neq j} \hat{Var}\left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)}{\sum_{i \neq j} \frac{s_{ij}^2}{s_{ii}s_{jj}}} = \frac{\sum_{i \neq j} \hat{Var}\left(\frac{s_{ij}}{\sqrt{1}}\right)}{\sum_{i \neq j} \frac{s_{ij}^2}{1}} = \frac{\sum_{i \neq j} \hat{Var}\left(s_{ij}\right)}{\sum_{i \neq j} s_{ij}^2}.$$

$\hat{\lambda}^{\star}$ increases if the ratio of samples and nodes decreases, i.e. if we have less samples or more nodes to rely on. The other reason for increase of the shrinkage intensity can be found in the denominator. If edges are deleted the correlations or rather covariances decrease and the ratio becomes greater.

In Figure 94 in the appendix the histograms of MaxDApC statistic for 300 and 100 samples and 100 samples per group without and with 50 additional nodes are drawn. No general differences in the distributions between 300 and 100 and 100 samples per group can be observed. The distribution of the maximal distances might be a bit shifted to right for the comparison of 300 and 100 samples which agrees with the slightly greater power of the test using the MaxDApC statistic. The reversed effect cannot be determined so easily for the mean distances that are equivalent to the MDApC statistic (cf. Figure 95 in the appendix).

The reason becomes obvious when we consider the effect of different numbers of nodes to condition on for estimating the partial correlations. The histograms of distances between partial correlations conditioned on the 11 original nodes of the RAF pathway and conditioned on the original 11 plus 50 additional nodes for 300 samples and 100 samples, where variance of the noise term is 1, can be seen in Figure 18. We take the previously simulated datasets with 300 samples in the first group and 50 additional nodes and the datasets with 100 samples in the first group without knockout (and 50 additional nodes) and estimate two partial correlations for every edge in every dataset.

First, we calculate the partial correlation conditioned on all other 59 variables, second, the additional nodes are removed and the partial correlation between the nodes of the original RAF network are computed. The histograms of these distances in Figure 18 show that for 100 samples a lot of distances are close to 0, but we observe also larger distances than for 300 samples. The many small distances can be explained by the fact that most partial correlations are estimated smaller in general compared to those estimated from 300 samples, because the covariances are shrinked more. Some partial

correlations that are estimated pretty large when conditioned on only 9 variables, will be more shrinked or rather the corresponding covariances, when we condition on 50 more nodes.

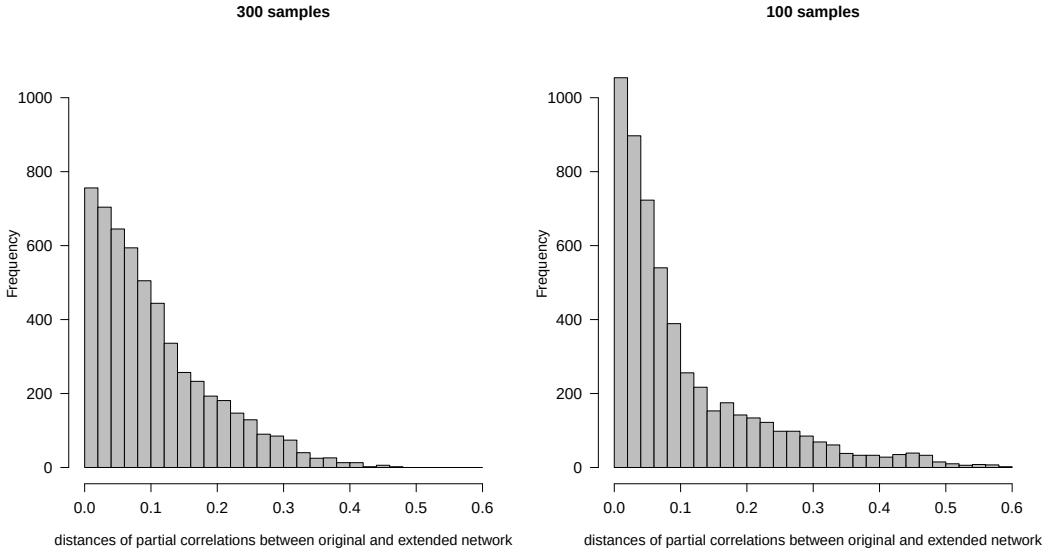

Figure 18: Histograms of distances between partial correlations conditioned on the 11 original nodes of the RAF pathway and conditioned on the original 11 plus 50 additional nodes for 300 samples (left) and 100 samples (right), where variance of the noise term is 1.

This reinforces the advantage in power for unbalanced sample sizes on the MaxDApC test. We see, the change of partial correlations is different for different numbers of samples, which increases the average differences for unequal sample sizes, also for permutated data. Hence, the tests based on averages of partial correlation distances loose power with growing number of nodes.

Estimation of partial correlations without shrinkage of the covariance matrix

Since the shrinkage of the covariance matrix seems to cause problems in several situations we perform the tests using MaxDApC, MDApC, and MDQpC again, but without

any shrinkage of the covariance matrix. Of course, now we need to restrict on scenarios where more samples per group than nodes are simulated. That means for setting with 50 samples in one or both groups the maximum number of additional nodes can be 20, if we simulate only 20 samples in at least one group we can just add 5 nodes to the 11 nodes of the RAF network structure. In addtion to the difference between partial correlations calculated with and without shrinkage of the covariance matrix of the three tests, we compare the results with the correlation based tests. The three tests based on the maximum, the mean absolute, and mean squared difference of (partial) correlations are selected, because they demontrated highest power under the alternative and hold the α-level in every setting.

In Table 19 the power differences of the tests using the maximum, the mean absolute, and mean squared difference between ordinary correlations and partial correlations with shrinkage of the covariance matrix are shown on the left and on the right side the distances between partial correlations with and without shrinkage. Distances are reported for all settings regarding numbers of additional nodes and variances of the noise term for 100 samples per group. The tables for the other considered sample sizes can be found in Tables 31 to 36 in the appendix. We see, in Table 19 the test using the maximal distance has considerably more power if ordinary correlations are used compared to partial correlations with shrinkage of the covariance matrix, especially with increasing numbers of additional nodes. If we compare the power between partial correlations with and without shrinkage, it is an advantage to shrink to covariances if no additional nodes or 50 additional nodes are considered with the maximal distance based test. For 5 to 20 additional nodes there are no considerable differences in the power.

Considering the mean absolute and mean squared difference the tests based on partial correlation with shrinkage of the covariance matrix have more power in each setting regarding number of additional nodes and variance of the noise term compared to the tests based on partial correlations without shrinkage. If many additional nodes are simulated the test using mean absolute distance with partial correlation with shrinkage has more power than the one using ordinary correlations, especially for moderate noise levels. But in general, we could rank the tests according to power, where ordinary correlation is better than partial correlation with shrinkage that in turn is better than partial correlation without shrinkage of the covariance matrix to recognize differences in two networks.

		Cor − pCor with shrinkage					pCor with − pCor without shrinkage				
	variance	number of additional nodes					number of additional nodes				
		0	5	10	20	50	0	5	10	20	50
MaxDA	0.01	0.10	0.53	0.71	0.76	0.58	0.22	-0.14	-0.19	-0.28	-0.09
	0.1	0.05	0.36	0.58	0.61	0.58	0.21	-0.08	-0.10	-0.26	-0.05
	0.5	0.00	0.34	0.57	0.66	0.63	0.29	0.03	-0.17	-0.10	0.09
	1	0.06	0.43	0.48	0.67	0.55	0.36	-0.04	-0.02	-0.02	0.15
	2	0.13	0.48	0.62	0.66	0.58	0.32	-0.02	-0.10	-0.04	0.09
	4	0.21	0.57	0.59	0.57	0.44	0.25	-0.12	0.04	-0.03	0.12
	8	0.32	0.46	0.56	0.48	0.33	0.25	0.00	-0.03	0.04	0.21
	16	0.25	0.37	0.33	0.34	0.10	0.18	0.05	0.05	0.03	0.17
MDA	0.01	0.08	0.08	0.13	0.16	-0.09	0.45	0.71	0.60	0.58	0.64
	0.1	0.07	-0.01	0.03	0.07	-0.11	0.48	0.57	0.72	0.51	0.51
	0.5	0.00	0.00	0.05	-0.01	-0.15	0.56	0.58	0.67	0.62	0.51
	1	0.00	0.08	0.17	0.12	-0.15	0.63	0.53	0.43	0.53	0.64
	2	0.25	0.26	0.30	0.16	-0.18	0.34	0.36	0.30	0.36	0.50
	4	0.30	0.37	0.30	0.06	-0.12	0.25	0.16	0.24	0.38	0.36
	8	0.40	0.41	0.18	0.12	-0.10	0.21	0.05	0.21	0.21	0.29
	16	0.23	0.16	0.06	0.02	-0.04	0.09	0.13	0.14	0.12	0.06
MDQ	0.01	0.08	0.12	0.23	0.31	0.07	0.34	0.59	0.48	0.44	0.65
	0.1	0.08	0.03	0.17	0.23	0.09	0.37	0.44	0.57	0.38	0.57
	0.5	0.00	0.02	0.16	0.23	0.07	0.45	0.47	0.48	0.45	0.56
	1	-0.03	0.11	0.23	0.31	0.05	0.61	0.46	0.38	0.44	0.64
	2	0.07	0.27	0.29	0.34	0.15	0.48	0.28	0.33	0.30	0.51
	4	0.20	0.39	0.48	0.23	0.08	0.37	0.19	0.21	0.34	0.36
	8	0.34	0.43	0.23	0.28	0.06	0.29	0.08	0.24	0.26	0.29
	16	0.25	0.17	0.16	0.11	0.01	0.16	0.20	0.20	0.17	0.07

Table 19: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 100 samples per group.

For the mean distances in the settings with unequal samples sizes and moderate noise levels, using partial correlations without shrinkage results in slightly higher power compared with shrinkage, but the tests using ordinary correlations is still more powerful.

An exception can be observed in case of small sample sizes of 20 or 50 samples per group. Here, the test using the mean absolute distance of partial correlations has more power levels than the same test based on ordinary correlations particularly for low noise, though the latter is still more powerful than the one that uses partial correlations without shrinkage of the covariance matrix.

Optimal distance for a test statistic to recognize differences in interaction networks

The questions which distance should be used to maximize the power is still to be addressed. Thereto, we record the maximum power over the ordinary correlation, partial correlation with and without shrinkage for every of the three statistics.
The differences between tests based on the three statistics optimized for the correlation examplified for 100 and 200 samples per group can be found in Tables 37 and 38 in the appendix. The results for the settings with other sample sizes look very similar and hence are not shown. The mean squared differences have more or approximate power (in the setting where only 20 samples per group are assumed) than the test based on mean absolute differences, particulary if many additional nodes are simulated and a high variance of the noise term is assumed. Notably, with many additional nodes and high variance the maximum distance is even more powerful than the mean squared distance. For moderate combinations of noise and numbers of additional nodes both statistics lead to similar results. However, the mean squared distance has advantages in terms of power for small sample sizes and low levels of variances of the noise terms.

4.4 Summary

Genes usually do not act alone but in groups or pathways. The identification and statistical inference of these pathways under certain conditions from gene expression data has become a recent research topic. Genes are treated as nodes and the interactions between the genes are represented as edges. Many approaches for network inference have been proposed. Relevance networks (Butte et al., 2000) derived from ordinary correlations of

genes are computationally inexpensive but they have two major disadvantages. First, the direction of the relationship can not be determined and second, we are not able to distinguish between direct and indirect interactions which means two genes are influenced by a third gene but have no direct relationship. Bayesian networks derived via Markov Chain Monte Carlo (MCMC) Simulations have the advantage to estimate posterior probabilities for directed egdes but they are computationally expensive. Another approach to estimate gene expression networks are Gaphical Gaussian Models (GGMs). Using partial correlations that can be easily obtained from the inverse of the covariance matrix direct interactions can be recognized and like relevance networks GGMs are computationally inexpensive. Only the direction of interactions remains unclear and we should keep in mind that a partial correlation depends on the other variables considered in the network.

If we assume nodes to be random variables and interactions between genes or nodes to be conditional independence structures we are able to extend the concept of conditional independence with the so called Markov properties (cf. Section 4.1.1) to a set of nodes and edges, i.e. a graph (cf. Section 4.1.1). Under the assumption that the genes or rather underlying random variables follow a multivariate normal distribution we speak of Graphical Gaussian Models. Lauritzen (1996, prop. 5.2, p. 129) shows that in this case two genes are independent conditional on another gene if and only if the conresponding entry of the concentration matrix, which is the inverse of the covariance matrix, is equal to zero. This is the link to the matrix of partial correlations respresenting the correlation of two variables given the values of other variables or phrased in a different way, the correlation of the residuals of the two variables fitted by a linear model with the other variables as independent variables. However, the covariance matrix can‘t be estimated if more observations than variables are considered which is often the case for gene expression measurements. A loophole for that purpose is shrinkage or biased estimation of the covariance matrix (Schäfer and Strimmer, 2005a, cf. section 4.1.3). Based on the theorem of Ledoit and Wolf (2003) the covariances are shrinked towards zero. The GGM estimated from these covariances outperforms GGM selection using Lasso regression (Meinshausen and Bühlmann, 2005) and other estimators for partial correlations that employ the pseudoinverse instead of the matrix inverse or that uses bootstrapping to obtain a variance reduced positive definite estimate of the covariance matrix (Schäfer and Strimmer, 2005b). Following Efron (2005) a mixture distribution for the observed partial correlations is assumed in order to compute the probability for

an existing edge, also reffered as local false discovery rate (cf. Section 4.1.4).
In this thesis we focus on the identification of differences in gene expression networks between two groups of patients or under two conditions. One can imagine that interactions between genes change with progression of a tumor disease, e.g. paths might collapse or activation of an oncogene might be increased through a certain signal transduction cascade.
To assess the differences of two networks or of one network under two conditions, we apply the mean absolute distance of partial correlations (Gill et al., 2010) and 13 novel statistics (cf. Section 4.1.5). Most of them are described in Lohr et al. (2010). Some statistics base on partial correlations, others on ordinary correlations. Some use the local false discovery rate to decide whether an edge is present or absent and some just depend on edges that are present in at least one network. To make it more robust ranks are considered for some statistics. All measures are used as test statistics in permutation tests (cf. Section 4.1.6), because they follow no known distribution.
But before we can test for differences in a network, a suitable set of genes that form a biological network or pathway and of which we can assume to be different between the two groups must be explored. Several databases on the internet, e.g. Gene Ontology, Reactome or KEGG provide predefined gene sets and pathway information from biological knowledge. Since we have no knowledge about differential interactions in a particular gene set, all gene sets need to be tested.
To avoid testing thousands of gene sets, we analyze gene signatures that are known to be differential between two groups of patients, though they might not build an interaction network. Therefore we perform a Gene Set Enrichment Analysis (GSEA, Mootha et al., 2003) for genes of differential signatures in predefined Gene Ontology gene sets. Enriched gene sets are afterwards tested for differences in interaction networks with our proposed statistics. The MammaPrint Genes (Tian et al., 2010) that are associated with prognosis are significantly enriched in 58 Gene Ontology gene sets. For 35 gene sets at least one test for differences in interaction networks is significant ($p < 0.05$). However, large discrepancies in the number of significantly different networks across the statistics are observed. The test depending on the maximum distance of partial correlations is not significant for any of the tested gene sets, while the two tests using the mean squared and absolute distances of ordinary correlations recognize 16 differential interaction networks.
A novel approach for the detection of differential networks is introduced in Section 4.2.2. The Gene Selection Algorithm for Differential Networks (DiNGS) for variable/gene se-

lection consists of five steps. These are exchangeable and may be adapted according to requirements. One basic version is described and analyzed for stability in the selection of the gene pair to built the (differential) network around.
To analyze properties in terms of type I errors and power, we perform an extended simulation study. Data has been generated based on the well-known structure of the RAF pathway, which consists of eleven phosphorilated proteins connected by 20 directed edges. Our tests base on partial or ordinary correlations that are only able to detect undirected edges. We need to test them on directed data, because of the biological rational that one protein or transcription factor influences another gene. Each node is modeled as linear combination of its parent nodes and an additional noise. The noise, the sample size of the groups that should be compared, the differences between the groups as well as the number of additional nodes that would not belong to the network, are varied. The additional nodes are generated to cause noise.
To check whether the proposed tests hold a given α-level, 1000 datasets without differences between the two assumed groups have been generated. Considering small sample sizes in at least one group, the tests using the local FDR to decide if an edge is present or absent, do not hold the given α-level.
Afterwards, networks with varying differences between two groups are generated to assess an estimate for the power of the proposed tests. In general, we can summarize that the power decreases with increasing number of additional "noise" nodes and higher variance of the noise term and increases with higher sample size. Furthermore, the power decreases significantly for the tests based on mean squared or absolute distances of partial correlations if the number of samples in one group is considerably higher than in the other one, compared to groups with smaller but equal sample sizes. Three reasons are suggested: First, the average might play a role, because the phenomenon is not observed for the maximum distance. Second, considering partial correlation may cause the effect, because tests using ordinary correlations are not affected, which lead to the third point. The shrinkage of covariances might cause the effect due to different shrinkage intensities caused by different sample sizes. Therefore, we applied the three tests with highest power - maximum distance, average of squared and absolute distances, once more without shrinkage of the covariance matrix. Of course, only situations with more observations than edges could be considered. A comparison with their counterparts with shrinked covariances and ordinary correlations lead to the following:
The tests using maximum or mean squared ordinary correlation have highest power in

all settings if the numbers of samples per group are equal and approximately twice as many samples per group than nodes are to be tested. The permutation test with the mean absolute distance of partial correlations (derived from shrinked covariances) as test statistic has slightly less but also high power. Assuming small variances and small sample sizes in conjuction with a small number of nodes in the network, the test using mean squared distance of partial correlations (with shrinkage of the covariance matrix) has the highest power. The latter can also be used for higher variances. In all other settings, it is advisable to use the permutation test with the maximum distance of ordinary correlations as test statistic. With increasing variance and numbers of nodes, the more superior is the maximum distance of ordinary correlations compared to the mean squared and absolute ordinary correlation and all other measures used as test statistics for the permutation tests.

5 Discussion and conclusions

High-dimensional gene expression data offers the opportunity to gain deeper insights into cancer biology which may help to develop novel therapies. The large amount of data may indeed be useful for that purpose, however, appropriate analysis strategies are required. Our goal was to improve statistical methods for extracting useful information about differences in gene expression with a focus on two topics - the validation of single genetic markers in multiple datasets and the detection of differential interaction networks among two groups of patients under two conditions.

The identification of differentially expressed genes between normal and tumor tissue, prognostic and predictive markers, from gene expression datasets is a major research topic. Due to the high number of measured genes, the chance of observing false positive findings is high. Usually a procedure to control the number or the proportion of false positive results is applied, but even after correcting for multiple testing we will obtain some false positive findings. Validation on other datasets will help to gain confidence in significant markers. By a strict adjustment on every considered dataset many interesting markers will not be recognized. Powerful standard approaches to combine estimators from different studies like the common meta-analysis (Whitehead, 2002) do not take the validation idea into account.
We proposed two strategies that trade adjustment for multiple testing in high-dimensional data off against validation of findings. Following the first, we screen for significant features in one dataset and afterwards a meta-analysis is performed for genes found to be interesting in the first step on other datasets to validate findings. Another strategy was called sequential validation strategy. Starting on one dataset, we test for significant features and all non-significant genes after adjustment are excluded from the next steps.

The procedure is repeated on the second dataset, followed by the third dataset until all datasets are tested and p-values were adjusted for an ever-decreasing number of features. To assess the characteristics, advantages and disadvantages of the proposed methods, we performed a simulation study and applied them to three real breast cancer ("Mainz", "Rotterdam" and "TRANSBIG") and nine non-small cell lung cancer datasets. The results were compared to those obtained by an ordinary meta-analysis.

Our two-step meta-analysis approach demonstrated its ability to identify additional prognostic genes in non-small cell lung cancer that would not have been recognized if all considered datasets were analyzed equally in an ordinary meta-analysis.

Applying the proposed sequential 3-step validation strategy, it seems sufficient for the elimination of all false positive features to restrict on only three datasets. The overlap of significant features of the latter strategy with an ordinary meta-analysis is even smaller when compared to the two-step meta-analysis. The outcome of our sequential validation strategy depends on the datasets used, the number of validation steps and the order of datasets. Testing on one and validating the results on two more datasets seems sufficient to exclude all false positive findings — at least in a simulation study. Although it seemed sufficient in the simulation studies, two validation steps might not be the optimal number to extract most signifcant but not false positive features since it depends on many factors.

Homogeneity of datasets is important since the highest number of significant genes after two validation steps is observed analyzing non-small cell lung cancer if datasets containing only patients with the same histological subtype are considered. Hence, testing all combinations and orders of datasets our approach offers a selection of datasets that are most homogeneous and therefore suitable for validating our findings.

In addition, we discovered that the sequential validation method enables us to draw conclusions about the quality in terms of noise and sample size of datasets in relation to each other. The breast cancer TRANSBIG cohort was identified to have the least quality since the sample size is larger than in the Mainz cohort, but less significant genes than in the Mainz and Rotterdam dataset are found. The least quality might arise from higher heterogeneity, since it is composed of samples from five European cancer centers. Zehetmayer and Posch (2012) proposed to conduct a small pilot-study first and validate the findings on a larger one. But our results argue against this strategy. We discovered that starting with the dataset with highest quality will result in increasing power.

The analysis of single genes allows only limited insights in biological processes that are disturbed in cancer. Considering interactions between genes in sub-networks seems to be more promising for that purpose.
We proposed the DiNGS (Differential Networks Gene Selection) Algorithm to detect differential interaction sub-networks. This algorithm can be used flexibly to build networks that, afterwards, can be tested explicitly for differences in the interaction structure. Since all of the five steps are exchangeble we might detect networks that are overall different between two groups or, we could extract a network with maximal differences in a small part, e.g. a path, in a graph. We are also free to build a larger network around the assumed differences or to restrict on the basic differential network. This is an advantageous feature, because genetic sub-networks cannot always be clearly distinguished from other sub-networks.
To test the hypothesis of differences in an interaction structure of a sub-network obtained by DiNGS or biological knowledge, we proposed a collection of measures and performed extensive simulation studies. Results were only shown for knockout of nodes "PKC" and "PIP2-PKC-PKA", while knockout of the other nodes led to similar findings.
We discovered that the permutation test with test statistic *MADPC* (mean absolute distance of partial correlations) proposed by Gill et al. (2010) has always less power than other statistics we proposed, e.g. maximum or mean squared distance of partial or ordinary correlations. Although Graphical Gaussian Models (GGMs) that base on partial correlations are known to have a better ability for network reconstruction than relevance networks based on ordinary correlations (Schäfer and Strimmer, 2005a), our tests using ordinary correlations have considerably more power.
The permutation tests using a local False Discovery Rate to decide whether an edge is present or absent have major disadvantages in terms of power and holding the α-level. A threshold of 0.2 as proposed by Efron (2005) might not be the best choice and could be adapted in future studies.
Unbalanced sample sizes between the two groups caused issues in most proposed permutation tests. Again, by using a test statistic with ordinary correlations instead of partial correlations we avoid this issue. Therefore, we argue for using a test based on ordinary correlation.

Although Schäfer and Strimmer (2005a) showed that GGMs using their shrinkage ap-

proach outperforms GGMs selection using Lasso regression (Meinshausen and Bühlmann, 2005) and other estimators for partial correlation that employ the pseudoinverse instead of matrix inverse or bootstrapping approaches, the concept of covariance shrinkage for differential network recognition should be reassessed in future work.
An extension on larger networks is computationally feasible for all proposed measures, but for these scenarios the properties of the tests will need to be investigated by additional simulation studies. Especially, when we analyze larger sub-networks a closer look into the structure will be necessary to explicitely find the differences. Therefore, an adaption of the testing procedure that tests at first for differential modular structures, then for differences in a sub-class of genes and finally for differential connectivity of single genes, proposed by Gill et al. (2010), might be a good approach to gain further insights into mechanisms responsible for cancer development or progression.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Alberts, B., Johnson, A., Walter, P., Lewis, J., Raff, M., and Keith Roberts, K., (2007). *Molecular Biology of the Cell*. 5. edition, Taylor & Francis.

[3] Alberg, A.J. and Samet, J.M. (2010). *Murray & Nadel's Textbook of Respiratory Medicine*. 5. edition, Saunders Elsevier.

[4] Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

[5] American Joint Committee on Cancer (2010). What is Cancer Staging? URL: *http://www.cancerstaging.org/mission/whatis.html*, retrieved 2010-10-05.

[6] American Cancer Society (2013). Cancer Facts and Figures 2013. Available online Exit Disclaimer, retrieved 2013-10-24.

[7] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Bioinformatics*, **25**, 25–9.

[8] Becker, N. and Wahrendorf, J. (1998). *Atlas of Cancer Mortality in the Federal Republic of Germany 1981–1990*. 3. edition, Springer.

[9] Benjamini,Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**(1), 289–300.

[10] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist*, **29**(4), 1165–1188.

[11] Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., Olson, J.A. Jr., Marks, J.R., Dressman, H.K., West, M., and Nevins, J.R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**(7074), 353–7.

[12] Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Journal of Clinical Pharmacy and Therapeutics*, **69**, 89–95.

[13] Bloom, H.J. and Richardson, W.W. (1957). Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, **11**(3), 359–377.

[14] Boes, T. (2007). *Auswirkungen der Low-Level-Analyse auf die Ergebnisse von Genexpressionsdaten der Firma Affymetrix*. Medizinische Fakultät der Universität Duisburg-Essen, Dissertation.

[15] Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

[16] Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., Berglund, A., Ekman, S., Bergqvist, M., Pontén, F., König, A., Fernandez, O., Karlsson, M., Helenius, G., Karlsson, C., Rahnenführer, J., Hengstler J.G., and Micke, P. (2013). Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis and tissue microarray validation. *Clinical Cancer Research*, **19**, 194–204.

[17] Brown, T.A., Yang, T.M., Zaitsevskaia, T., Xia, Y., Dunn, C.A., Sigle, R.O., Knudsen, B., and Carter, W.G. (2004). Adhesion or plasmin regulates tyrosine phosphorylation of a novel membrane glycoprotein p80/gp140/CUB domain-containing protein 1 in epithelia. *Journal of Biological Chemistry*, **279**, 14772–14783.

[18] Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R., and Kohane, I.S. (2000). Discovering functional relationships between RNA expression and

chemotherapeutic susceptibility using relevance networks. *Proceedings of The National Academy of Sciences*, **97**, 12182–12186.

[19] Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A.M., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., and Piccart, M.J. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst*, 98(17), 1183–92.

[20] Cancer Genome Atlas Research Network. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

[21] Carmona, R.H. (2006). The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General, U.S. Department of Health and Human Services, URL: *http://www.surgeongeneral.gov/library*, retrieved 2006-06-27.

[22] Choi, J.K., Yu, U., Yoo, O. J. and Kim, S. (2005). Differential co-expression analysis using microarray data and its application to human cancer. *Oxford Journal Bioinformatics*, **21**, 4348–4355.

[23] Clifford, P. (1990). Markov random fields in statistics. *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, Oxford University Press, 19–32.

[24] Clopper, C. and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

[25] Cooper, W.A., O'toole, S., Boyer, M., Horvath, L., and Mahar ,A. (2011). What's new in non-small cell lung cancer for pathologists: the importance of accurate subtyping, EGFR mutations and ALK rearrangements. *Pathology*, **43**,103–115.

[26] Coussens, L., Yang-Feng, T.L., Liao, Y.C., Chen, E., Gray, A., McGrath, J., Seeburg, P.H., Libermann, T.A., Schlessinger, J., and Francke, U.

(1985). Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science*, **230**(4730), 1132–9.

[27] Croce, C.M. (2008). Oncogenes and cancer. *The New England Journal of Medicine*, **358**, 502–11.

[28] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2008). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–7.

[29] Cyris, C. (2011). *Auswahl von Genen zur Schätzung differentieller Netzwerke*. Fakultät Statistik der Technischen Universität Dortmund, Bachelor thesis.

[30] Daniels, M.J. and Kass, R.E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, **57**, 1173–1184.

[31] Dempster, A. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

[32] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177–188.

[33] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A.L., Klijn, J.G., Foekens, J.A., Cardoso, F., Piccart, M.J., Buyse, M., and Sotiriou, C. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*, 13(11), 3207

[34] Deutsches Krebsforschungszentrum (DKFZ), (2013). URL: *http://www.dkfz.de/de/presse/download/Krebs_Lunge.pdf*, retrieved 2013-09-13.

[35] Dougherty, M.K., Müller, J., Ritt, D.A., Zhou, M., Zhou, X.Z., Copeland, T.D., Conrads, T.P., Veenstra, T.D., Lu, K.P., and Morrison, D.K. (2005). Regulation of Raf-1 by direct feedback phosphorylation. *Molecular Cell*, **17**, 215–24.

[36] Dudoit, S., Popper Shaffer, J., and Boldrick J.C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, **18**(1), 71–103.

[37] Dudoit, S. and van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer.

[38] Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–10.

[39] Edwards, D. (2000). *Introduction to Graphical Modelling*, 2. edition, Springer.

[40] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choise of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

[41] Efron, B. (2005). Local false discovery rates. Preprint, *Dept. of Statistics*, Stanford University.

[42] Eheman, C.R., Shaw, K.M., Ryerson, A.B., Miller, J.W., Ajani, U.A., and White, M.C. (2009). The changing incidence of in situ and invasive ductal and lobular breast carcinomas: United States, 1999–2004. *Cancer Epidemiology, Biomarkers & Prevention*, **18**(6):1763–9.

[43] Elston, C.W and Ellis, I.O. (1991). Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. *Histopathology*, **19**, 403–410.

[44] Evidence-Based Medicine Working Group (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*, **268**(17), 2420–5.

[45] Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers,C., and Parkin D.D. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*, **127**, 2893–2917.

[46] Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

[47] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Oxford Journal Bioinformatics*, **9**, 432–441.

[48] Fritz, C.C., Zapp, M.L., and Green, M.R. (1995). A human nucleoporin-like protein that specifically interacts with HIV Rev. *Nature*, **376**, 530–533.

[49] de la Fuente, A. (2010). From differential expression to differential networking — identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, **26**(7), 326–33.

[50] Fujikoshi, Y., Ulyanov, V.V., and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. John Wiley & Sons, Inc.

[51] Gambardella, G., Moretti, M. N., de Cegli, R., Cardone, L., Peron, A. and di Bernardo, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Oxford Journal Bioinformatics*, **29**, 1776–1785.

[52] Gautier, L., Cope, L., Bolstad, B.M., and Irizarry R.A. (2004) affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3), 307–15.

[53] Genschel, U. and Becker, C. (2005). *Schließende Statistik: Grundlegende Methoden*. Springer.

[54] Gilbert, S.F. (2003). *Developmental Biology*. 7. edition, Sinauer Associates.

[55] Gillis, J. and Pavlidis, P. (2009). A methodology for the analysis of differential co-expression across the human lifespan. *BMC Bioinformatics*, **10**, 306.

[56] Glass, G.V. (1976). Primary, Secondary and Meta-Analysis of Research. *Educational Researcher*, 5, 3–8.

[57] Grzegorczyk, M. and Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.

[58] Goldstraw, P., Crowley, J., Chansky, K., Giroux, D.J., Groome, P.A., Rami-Porta, R., Postmus, P.E., Rusch, V., and Sobin, L. (2007). The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol*, **2**, 706–714.

[59] Haar A. (1909). *Zur Theorie der orthogonalen Funktionensysteme.* Mathematische Annalen.

[60] Hammersley, J.M. and Clifford, P. (1971). *Markov fields on finite graphs and lattices.* Unpublished.

[61] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer.

[62] Hilger, R.A., Scheulen, M.E., and Strumberg, D. (2002). The Ras-Raf-MEK-ERK pathway in the treatment of cancer. *Onkologie*, **25**, 511–8.

[63] Hochberg, Y. and Tamhane, A.C. (1987). *Multiple Comparison Procedures*, John Wiley & Sons, Inc.

[64] Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55–67.

[65] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

[66] Horn, L., Pao., W., and Johnson, D.H. (2012). "Chapter 89". In Longo, D.L., Kasper, D.L., Jameson, J.L., Fauci, A.S., Hauser, S.L., and Loscalzo, J. *Harrison's Principles of Internal Medicine.* 18. Edition. McGraw-Hill.

[67] Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistic Society B*, **15**, 193–232.

[68] Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J.A., Hoogsteden, H.C, Grosveld, F.,and Philipsen, S. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*, **5**(4), e10312.

[69] Ikeda, J.I., Morii, E., Kimura, H., Tomita, Y., Takakuwa, T., Hasegawa, J.I., Kim Y.K., Miyoshi Y., Noguchi S., Nishida T., and Aozasa K. (2006). Epigenetic regulation of the expression of the novel stem cell marker CDCP1 in cancer cells. *Journal of Pathology*, **210**, 75–84.

[70] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*.

[71] Ismail R.S., Baldwin R.L., Fang J., Browning D., Karlan B.Y., Gasson J.C., Chang D.D. (2000). Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer Research*, **60**, 6744–9.

[72] Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J.E., Liu, E.T., Bergh, J., Kuznetsov, V.A., and Miller, L.D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, **66**, 10292–301.

[73] Jacob L., Neuvial P., and Dudoit, S. (2012). Gains in Power from Structured Two-Sample Tests of Means on Graphs. *The Annals of Statistics*, **6**, 561–600.

[74] Jemal, A., Tiwari,R.C., Murray, T., Ghafoor, A., Samuels, A., Ward, E., Feuer, E.J., and Thun, M.J. (2004). Cancer statistics 2004. *CA Cancer J Clin*, **54**(1), 8–29.

[75] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, **61**, 69–90.

[76] Jett, J.R., Schild, S.E., Keith, R.L., and Kesler K.A. (2007). Treatment of non-small cell lung cancer, stage IIIB: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, **132**, 266S–276S.

[77] Juric, D., Lacayo, N.J., Ramsey, M.C., Racevskis, J., Wiernik, P.H., Rowe, J.M., Goldstone, A.H., O'Dwyer, P.J., Paietta, E., and Sikic, B.I. (2007). Differential gene expression patterns and interaction networks in BCR-ABL-positive and -negative adult acute lymphoblastic leukemias. *Journal of Clinical Oncology*, **25**, 1341–9.

[78] Kaatsch, P., Spix, C., Katalinic, A., Hentschel, S., Baras, N., Barnes, B., Bertz, J., Dahm, S., Haberland J., Kraywinkel K., Laudi, A., and Wolf, U. (2012). Krebs in Deutschland 2007/2008, Häufigkeiten und Trends. *Eine gemeinsame Veröffentlichung des Robert Koch-Instituts und der Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V.*, 8. edition.

[79] Kenehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**(1), 27–30.

[80] Kaufman, L. and Rousseeuw, P.J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*, 1. edition, Wiley-Interscience.

[81] Kindermann, R. and Snell, J.L. (1980). *Markov Random Fields and Their Applications*, American Mathematical Society.

[82] Klein, J.P. and Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2. edition, Springer.

[83] Knudson, A.G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, **1**, 157–62.

[84] Kostka, D. and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Oxford Journal Bioinformatics*, **20**, i194–i199.

[85] Krämer, N., Schäfer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics*, **10**, 384.

[86] Langer, C.J., Besse, B., Gualberto, A., Brambilla, E., and Soria, J.C (2010). The evolving role of histology in the management of advanced non-small-cell lung cancer. *J Clin Oncol*, **28**, 5311–5320.

[87] Lauritzen, S.L. (1996). *Graphical Models*, Oxford University Press.

[88] Law, M.L., Kao, F.T., Wei, Q., Hartz, J.A., Greene, G.L., Zarucki-Schulz, T., Conneely, O.M., Jones, C., Puck, T.T., and O'Malley, B.W. (1987). The progesterone receptor gene maps to human chromosome band 11q13, the site of the mammary oncogene int-2". *Proc. Natl. Acad. Sci. U.S.A.*, **84**(9), 2877–81.

[89] Lazaridis, E. N., Sinibaldi, D., Bloom, G., Mane, S., and Jove, R. (2002). A simplemethod to improve probe set estimates from oligonucleotide arrays. *Math Biosci*, **176**(1), 53–58.

[90] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, **10**, 603–621.

[91] Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*, 3. edition, Springer.

[92] Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Oxford Journal Bioinformatics*, **24**, 1175–1182.

[93] Lipshutz,R., Fodor,S., Gingeras,T., and Lockhart D. (1999). High density synthetic oligonucleotide arrays. *Nature Genet*, **21**, 20–24.

[94] Liu, R., Wang, X., Chen, G.Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., and Clarke, M.F. (2007). The prognostic role of a gene signature from tumorigenic breast-cancer cells. *The New England Journal of Medicine*, **356**, 217–26.

[95] Liu, D.C., Yang, Z.L., and Jiang, S. (2011). Identification of PEG10 and TSG101 as carcinogenesis, progression, and poor-prognosis related biomarkers for gallbladder adenocarcinoma. *Pathology Oncology Research*, **17**, 859–66.

[96] Lohr, M., Godoy, P., Hengstler, J.G., Rahnenführer, J., and Grzegorczyk, M. (2010). Extracting differential regulatory sub-networks from genome-wide microarray expression data. *Proceedings of the Seventh International Workshop on Computational Systems Biology*, 63–66.

[97] Lohr, M., Köllmann, C., Freis, E., Hellwig, B., Hengstler, J.G., Ickstadt, K., and Rahnenführer, J. (2012). Optimal Strategies for Sequential Validation of Significant Features from High-Dimensional Genomic Data. *Journal of Toxicology and Environmental Health*, **75**, 8(10), 447–460.

[98] Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J.A., Klijn, J.G., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M.J., and Sotiriou, C. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol*, **25**(10), 1239–46.

[99] Lum, K.L., Polansky, J.R., Jackler, R.K., and Glantz, S.A. (2008). Signed, Sealed and Delivered: "Big Tobacco" in Hollywood, 1927–1951. *Tobacco Control*, **17**, 313–323.

[100] Lux, A., Beil, C., Majety, M., Barron, S., Gallione, C.J., Kuhn, H.-M., Berg, J.N., Kioschis, P., Marchuk, D.A., Hafner, M. (2005). Human retroviral gag- and gag-pol-like proteins interact with the transforming growth factor-beta receptor activin receptor-like kinase 1. *Journal of Biological Chemistry*, **280**, 8482–8493.

[101] Mansmann, U., Schmidberger, M., Strobl, R., and Jurinovic, V. (2010). *Indirect Comparison of Interaction Graphs — Statistical Modelling and Regression Structures*. Physica-Verlag HD.

[102] McCubrey, J.A., Steelman, L.S., Chappell, W.H., Abrams, S.L., Wong, E.W., Chang, F., Lehmann, B., Terrian, D.M., Milella, M., Tafuri, A., Stivala, F., Libra, M., Basecke, J., Evangelisti, C., Martelli, A.M., and Franklin, R.A. (2007). Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. *Biochimica et Biophysica Acta*, **1773**, 1263–1284.

[103] Meinshausen, N. and Bühlmann, P. (2005). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, **34**, 1436–1462.

[104] Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., and Stein, L. (2012). Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. *Cancers*, **4**, 1180–1211.

[105] Miller, R.A., Galecki, A., and Shmookler-Reis, R.J. (2001). Interpretation, design, and analysis of gene array expression expression experiments. *J Gerontol A Biol*, **56**, B52–B57.

[106] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C. (2003). PGC-1 -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267–273.

[107] Mosca E., Alfieri R., Merelli I., Viti F., Calabria A., and Milanesi L. (2010). A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, **3**, 76.

[108] National Cancer Institute (2013). SEER Cancer Statistics Review. URL: *http://seer.cancer.gov/archive/csr/1975_2010/*, retrieved 2013-05-12.

[109] National Cancer Institute (2013). Lifetime Risk (Percent) of Being Diagnosed with Cancer by Site and Race/Ethnicity: Males, 18 SEER Areas, 2008–2010 (Table 1.16) and Females, 18 SEER Areas, 2008–2010 (Table 1.17). SEER Cancer Statistics Review. URL: *http://seer.cancer.gov/archive/csr/1975_2010/results_merged /topic_lifetime_risk_diagnosis.pdf*, retrieved 2013-05-12.

[110] National Cancer Institute (2013). Lifetime Risk (Percent) of Dying from Cancer by Site and Race/Ethnicity: Males, Total US, 2008–2010 (Table 1.19) and Females, Total US, 2008–2010 (Table 1.20). SEER Cancer Statistics Review. URL:

http://seer.cancer.gov/archive/csr/1975_2010/results_merged/topic_lifetime_risk_death.pdf, retrieved 2013-06-12.

[111] Netzer, T. (2013). *Vorhersage der Überlebenswahrscheinlichkeit für Patientenuntergruppen mit hochdimensionalen Daten am Beispiel zweier Lungenkrebskohorten*. Fakultät Statistik der Technischen Universität Dortmund, Dissertation.

[112] Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S., Watanabe, S., Sakamoto, H., Kumamoto, K., Takenoshita, S., Gotoh, N., Mizuno, H., Sarai, A., Kawano, S., Yamaguchi, R., Miyano, S., and Yokota, J. (2012). Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*, **72**(1), 100–11.

[113] Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., and Carbone, P.P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol*, **5**(6), 649–55.

[114] O'Reilly, K.M., Mclaughlin, A.M., Beckett, W.S., and Sime, P.J. (2007). Asbestos-related lung disease. *American Family Physician*, **75**(5), 683–688.

[115] Orton, R.J., Sturm, O.E., Vyshemirsky, V., Calder, M., Gilbert, D.R., and Kolch, W. (2005). Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. *The Biochemical Journal*, **392**, 249–61.

[116] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

[117] Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J.M., Macdonald, J., Thomas, D., Moskaluk, C., Wang, Y., and Beer, D.G. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res*, **66**(15), 7466–72.

[118] R Core Development Team (2013). R: A Language and Environment for Statistical Computing, Version 2.15.1, ISBN: 3-900051-07-0, URL: *http://www.R-project.org*.

[119] Révillion, F., Pawlowski, V., Hornez, L., and Peyrat, J.P. (2000). Glyceraldehyde-3-phosphate dehydrogenase gene expression in human breast cancer. *European Journal of Cancer*, **36**, 1038–42.

[120] Rosti, G., Bevilacqua, G., Bidoli, P., Portalone, L., Santo, A., and Genestrti G. (2006). Small cell lung cancer. *Ann Oncol*, **17**, ii5–10.

[121] Sachs, K., Perez, O., Peèr, D.A., and Nolan, G.P. (2005). Protein-Signaling networks derived from multiparameter single-cell data. *Science*, **208**, 523–529.

[122] Said, H.M., Polat, B., Hagemann, C., Anacker, J., Flentje, M., and Vordermark, D. (2009). Absence of GAPDH regulation in tumor-cells of different origin under hypoxic conditions in - vitro. *BMC Research Notes*, **2**, 8.

[123] Schäfer, J., and Strimmer, K. (2005a). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**(1), Art. 32.

[124] Schäfer, J. and Strimmer, K. (2005b). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

[125] Scherl-Mostageer, M., Sommergruber, W., Abseher, R., Hauptmann, R., Ambros, P., and Schweifer, N. (2001). Identification of a novel gene, CDCP1, overexpressed in human colorectal cancer. *Oncogene*, **20**, 4402–4408.

[126] Schmidt, M., Böhm, D., von Törne, Ch., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J.G., Kölbl, H., and Gehrmann, H. (2008). The Homural Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer. *Clinical Cancer Research*, **68**, 5405–5413.

[127] Schuhmacher, M. and Schulgen, G. (2006). *Methodik klinischer Studien*. 2. edition, Springer.

[128] Schulz, W.A. (2005). *Molecular Biology of Human Cancers — An Advanced Student's Textbook*, Springer.

[129] Schwarz, G.E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

[130] Scott, W.J., Howington, J., Feigenberg, S., Movsas, B., and Pisters, K. (2007). Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, **132**, 234S–242S.

[131] Sequist, L.V., Bell, D.W., Lynch, T.J., and Haber, D.A. (2007). Molecular predictors of response to epidermal growth factorreceptor antagonists in non-small-cell lung cancer. *J Clin Oncol*, **25**, 587–595.

[132] Shedden, K., Taylor, J.M., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., Misek, D.E., Chang, A.C., Zhu, C.Q., Strumpf, D., Hanash, S., Shepherd, F.A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V.E., Meyerson, M., Kuick, R., Dobbin, K.K., Lively, T., Jacobson, and J.W., Beer, D.G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med*, **14**(8), 822–7.

[133] Sirover, M.A. (1999). New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochimica et Biophysica Acta*, **1432**, 159–84.

[134] Sobin, L.H. and Compton, C.C. (2010). TNM seventh edition: what's new, what's changed: communication from the International Union Against Cancer and the American Joint Committee on Cancer. *Cancer*, 116, 5336–5339.

[135] Socinski, M.A., Crowell, R., Hensing, T.E., Langer, C.J., Lilenbaum, R., Sandler, A.B., and Morris, D. (2007). Treatment of non-small cell lung cancer, stage IV: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, **132**, 277S–289S.

[136] Spiro, S.G., Gould, M.K., and Colice, G.L., (2007). Initial evaluation of the patient with lung cancer: symptoms, signs, laboratory tests, and paraneoplastic syndromes: ACCP evidenced-based clinical practice guidelines (2nd edition). *Chest*, **132**, 149S–160S.

[137] Stangl D. and Berry D.A. (2000). *Meta Analysis in Medicine and Health Policy*. CRC Press.

[138] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, **6c4**, 479–498.

[139] Subramanian, J. and Govindan, R. (2007). Lung cancer in never smokers: a review. *Journal of Clinical Oncology*, **25**(5), 561–570.

[140] Sutton, A.J., Jones, D.R., Abrams, K.R., Sheldon, T.A., and Song, F. (2000). *Methods for Meta-analysis in Medical Research*. Wiley.

[141] Tenenhaus, A., Guillemot, V., Gidrol, X., and Frouin, V. (2010). Gene Association Networks from Microarray Data Using a Regularized Estimation of Partial Correlation Based on PLS Regression. *Computational Biology and Bioinformatics*, **7**, 251–262.

[142] Tian, S., Roepman, P., van't Veer, L.J., Bernards, R., de Snoo, F., and Glas, A.M. (2010). Biological Functions of the Genes in the Mammaprint Breast Cancer Profile Reflect the Hallmarks of Cancer. *Biomark Insights*, **5**, 129–138.

[143] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, **58**, 267–288.

[144] Travis, W.D. (2011). Pathology of lung cancer. *Clinics in Chest Medicine*, **32**, 669–692.

[145] Tsim, S., O'Dowd, C.A., Milroy, R., and Davidson, S. (2010). Staging of non-small cell lung cancer (NSCLC): a review. *Respir Med*, **104**, 1767–1774.

[146] Tsuji, K., Yasui, K., Gen, Y., Endo, M., Dohi, O., Zen, K., Mitsuyoshi, H., Minami, M., Itoh, Y., Taniwaki, M., Tanaka, S., Arii, S., Okanoue, T.,

and Yoshikawa, T. (2011). PEG10 is a probable target for the amplification at 7q21 detected in hepatocellular carcinoma. *Cancer Genetics and Cytogenetics*, **198**, 118–25.

[147] Toutenburg, H. (2009). *Deskriptive Statistik: Eine Einführung mit Übungsaufgaben und Beispielen mit SPSS*, 7. edition, Springer.

[148] van 't Veer L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–6.

[149] Victor, A. and Hommel, G. (2007). Combining adaptive designs with control of the false discovery rate — a generalized definition for a global p-value. *Biometrical Journal*, **49**, 94–106.

[150] Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, **54**(3), 426–482.

[151] Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D., and Foekens, J.A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**(9460), 671–9.

[152] Wang, N., Zhou, F., Xiong,H., Du, S., Ma, J., Okai, I., Wang, J., Suo, J., Hao, L., Song, Y., Hu, J., and Shao, S. (2012). Screening and identification of distant metastasis-related differentially expressed genes in human squamous cell lung carcinoma. *Anat Rec (Hoboken)*, **295**(5), 748–57.

[153] Werhli, A.V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **22**, 2523–2531.

[154] WHO World Health Organization (2003). *Tumours of the Breast and Female Genital Organs*, Oxford University Press.

[155] Windgassen, D. (2011). *Stabilitätsanalyse bei der Schätzung von differentiellen genetischen Netzwerken*. Fakultät Statistik der Technischen Universität Dortmund, Bachelor thesis.

[156] Whitehead, A. (2002). *Meta-Analysis of Controlled Clinical Trials*, John Wiley & Sons, Inc.

[157] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, Inc.

[158] Xie, Y., Xiao, G., Coombes, K.R., Behrens, C., Solis, L.M., Raso, G., Girard, L., Erickson, H.S., Roth, J., Heymach, J.V., Moran, C., Danenberg, K., Minna, J.D., and Wistuba, I.I. (2011). Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin Cancer Res*, **17**(17), 5705–14.

[159] Yamabuki, T., Takano, A., Hayama, S., Ishikawa, N., Kato, T., Miyamoto, M., Ito, T., Ito, H., Miyagi, Y., Nakayama, H., Fujita, M., Hosokawa, M., Tsuchiya, E., Kohno, N., Kondo, S., NakamuraY., and Daigo, Y. (2007). Dikkopf-1 as a Novel Serologic and Prognostic Biomarker for Lung and Esophageal Carcinomas. *Cancer Research*, **67**, 2517.

[160] Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, **1**, 217–235.

[161] Zehetmayer, S. and Posch, M. (2012). False discovery rate control in two-stage designs. *BMC Bioinformatics*, **13**, 81.

[162] Zhao, S.D., Cai, T.T., and Li, H. (2014). Direct estimation of differential network. *Oxford Journal Biometrika*, **101**, 253–268.

[163] Zheng, B., Fiumara, P., Li, Y.V., Georgakis, G., Snell, V., Younes, M., Vauthey, J.N., Carbone, A., and Younes, A. (2003). MEK/ERK pathway is aberrantly active in Hodgkin disease: a signaling pathway shared by CD30, CD40, and RANK that regulates cell proliferation and survival. *Blood*, **103**, 1019–1027.

[164] Zhu, C.Q., Ding, K., Strumpf, D., Weir, B.A., Meyerson, M., Pennell, N., Thomas, R.K., Naoki, K., Ladd-Acosta, C., Liu, N., Pintilie, M., Der, S., Seymour, L., Jurisica, I., Shepherd, F.A., and Tsao, M.S. (2010). Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol*, **28**(29), 4417–24.

Acknowledgements

First of all, I want to thank my primary supervisor Prof. Dr. Jörg Rahnenführer. Already during my Diploma studies he encouraged my enthusiams for applied biostatistics and the analysis of gene expression data with his joy and everlasting positive atmosphere. I thank him for countless hours of discussing and giving me food for though. He was an example for me and without him I would probably not working in the position where I am now. Even after I left the TU Dortund before I finished my dissertation, Jörg Rahnenführer continually motivated me to complete my thesis.
A big thank goes to Dr. Marco Grzegorczyk. All I learnt about (Baysian) network inference, I got from him. He gave input to a lot of issues concerning netwoks in my thesis.
My special thanks are also due JProf. Dr. Uwe Ligges who agrees to co-advise this work at short-notice after Marco Grzegorczyk left the TU Dortmund. He teached my lots of tricks in R and computer stuff, and always had a delicious coffee for me.
Most of my research was supported my the German Research Foundation (DFG, Grant RA 870/5–1). I am also thankful for support by the IfADo (Leibniz Research Center for Working Environment and Human Factors) and especially Prof. Dr. Jan Hengstler in the first months after my Diploma thesis.
I really appreciated my time at the TU Dortmund with all the nice collegues. Many friendships have been developed during the years. Thank you all for the great time I had in the coffee breaks, for lunch and on "wednesday meetings".
Last but not least, I would like to thank my family for everlasting love, support and confidence in me. Andreas, thank you for giving me the freedom to complete my PhD studies. Timo, you are the best brother in the world, I am so proud of you. Thank you, Muugi. Everything I am, I owe you.

Declaration

I declare that this thesis is written by myself and that I exclusively used the indicated literature and resources. The thoughts taken directly or indirectly from external sources are proper marked as such.

Appendix

Additional tables

	Description
Primary tumour (T)	
T1	Tumour =3 cm in diameter surrounded by lung or visceral pleura, without invasion more proximal than lobar bronchus.
T1a	Tumour =2 cm in diameter
T1b	Tumour >2 cm but =3 cm in diameter
T2	Tumour >3 cm but =7 cm in diameter or tumour with: -Involvement of the main bronchus =2 cm distal to the carina -Invasion of visceral pleura -Associated atelectasis or obstructive pneumonitis that does not involve the entire lung.
T2a	Tumour =5 cm in diameter
T2b	Tumour >5 cm but =7 cm in diameter
T3	Tumour >3 cm but =7 cm in diameter or tumour with: -Direct invasion of the chest wall, diaphragm, phrenic nerve -Direct invasion of the mediastinal pleura or parietal pericardium -Associated atelectasis or obstructive pneumonitis that involves the entire lung. -Tumour within the main bronchus < 2 cm to the carina, without involvement of the carina. -Satellite tumour nodule(s) in the same lobe.
T4	Tumour of any size with: -Invasion of mediastinum -Invasion of heart or great vessels -Invasion of trachea, oesophagus, or recurrent laryngeal nerve -Invasion of a vertebral body or carina -Separate tumour nodules in a different ipsilateral lobe.
Regional lymph node (N)	
N0	No regional lymph node metastasis
N1	Involvement of ipsilateral hilar or peribronchial nodes
N2	Involvement of ipsilateral mediastinal or subcarinal nodes
N3	Involvement of contralateral mediastinal or hilar nodes, or ipsilateral/contralateral scalene or supraclavicular nodes.
Distant metastasis (M)	
M0	No distant metastasis
M1	Distant metastasis present
M1a	Separate tumour nodule(s) in a contralateral lobe or tumour with pleural nodules or malignant pleural/pericardial effusion
M1b	Distant metastasis

Table 20: Definition of TNM (7. edition, reproducted from Goldstraw et al., 2007).

	$\sigma_{hq} = 0.7, \sigma_{mq} = 0.8, \sigma_{lq} = 1.8$					
	Holm			FDR		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $hq \to mq \to lq$	58 / 0	58 / 0	22 / 0	92 / 5	92 / 0	74 / 0
2: $hq \to lq \to mq$	58 / 0	22 / 0	22 / 0	92 / 5	73 / 0	73 / 0
3: $mq \to hq \to lq$	45 / 0	45 / 0	20 / 0	85 / 4	85 / 0	69 / 0
4: $mq \to lq \to hq$	45 / 0	20 / 0	20 / 0	85 / 4	69 / 0	69 / 0
5: $lq \to hq \to mq$	1 / 0	1 / 0	1 / 0	4 / 0	4 / 0	4 / 0
6: $lq \to mq \to hq$	1 / 0	1 / 0	1 / 0	4 / 0	4 / 0	4 / 0

Table 21: Median true positives/false positives in each step of the six adjustment sequences for simulated gene expression data.

	$\sigma_{hq} = 0.8, \sigma_{mq} = 1.1, \sigma_{lq} = 1.5$					
	FDR/Holm/Holm			Holm/FDR/FDR		
sequence	1st step	2nd step	3rd step	1st step	2nd step	3rd step
1: $hq \to mq \to lq$	85 / 4	74 / 0	39 / 0	45 / 0	45 / 0	42 / 0
2: $hq \to lq \to mq$	85 / 4	40 / 0	39 / 0	45 / 0	42 / 0	42 / 0
3: $mq \to hq \to lq$	52 / 2	52 / 0	32 / 0	17 / 0	17 / 0	16 / 0
4: $mq \to lq \to hq$	52 / 2	31 / 0	31 / 0	17 / 0	16 / 0	16 / 0
5: $lq \to hq \to mq$	15 / 1	15 / 0	15 / 0	4 / 0	4 / 0	4 / 0
6: $lq \to mq \to hq$	15 / 1	15 / 0	15 / 0	4 / 0	4 / 0	4 / 0

Table 22: Median true positives/false positives in each step of the six adjustment sequences for simulated gene expression data.

sequence					1st step	2nd step	3rd step
Jacob	→	GSE31547	→	GSE31210	258	0	0
Jacob	→	GSE31210	→	GSE31547	258	120	18
GSE31547	→	Jacob	→	GSE31210	0	0	0
GSE31547	→	GSE31210	→	Jacob	0	0	0
GSE31210	→	Jacob	→	GSE31547	4355	241	21
GSE31210	→	GSE31547	→	Jacob	4355	0	0

Table 23: Numbers of significant probe sets in each step for the six posiible validation sequences (using the Benjamini-Hochberg procedure for significance adjustment to restrict the FDR to 5%) for the three lung cancer datasets that afford the highest numbers of significant probe sets after three adjustment steps.

sequence					1st step	2nd step	3rd step
Jacob	→	GSE3141	→	GSE31547	258	19	5
Jacob	→	GSE3141	→	GSE31210	258	19	11
Jacob	→	GSE31210	→	GSE31547	258	120	18
Jacob	→	GSE31210	→	GSE3141	258	120	13
GSE31210	→	Jacob	→	GSE31547	4355	241	21
GSE31210	→	Jacob	→	GSE3141	4355	241	12

Table 24: Numbers of significant probe sets in each step for the validation sequences (using the Benjamini-Hochberg procedure for significance adjustment to restrict the FDR to 5%) that consider at least one significant probe set after three adjustment steps on the lung cancer datasets containing the adenocarcinoma patients.

probe set	symbol	gene name
201037_at	PFKP	phosphofructokinase, platelet
202616_s_at	MECP2	methyl CpG binding protein 2 (Rett syndrome)
204385_at	KYNU	kynureninase
205839_s_at	BZRAP1	benzodiazapine receptor (peripheral) associated protein 1
207165_at	HMMR	hyaluronan-mediated motility receptor (RHAMM)
214710_s_at	CCNB1	cyclin B1
218092_s_at	AGFG1	ArfGAP with FG repeats 1

Table 25: Probe sets with gene symbol and gene name that are significant in the combined meta-analysis for all patients when Uppsala is used as basic dataset as well if dataset GSE31210 is used for preselection.

number of nodes		network 1 present	network 1 absent
network 2	present	15	0
	absent	5	35

Table 26: Theoretical contingency table for simulated networks without any additional nodes. Value of the χ^2 statistic with continuity correction is $T_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = 32.4115$.

number of nodes		network 1 present	network 1 absent
network 2	present	15	0
	absent	5	100

Table 27: Theoretical contingency table for simulated networks with 5 additional nodes. Value of the χ^2 statistic with continuity correction is $T_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = 78.9943$.

number of nodes		network 1 present	network 1 absent
network 2	present	15	0
	absent	5	190

Table 28: Theoretical contingency table for simulated networks with 10 additional nodes. Value of the χ^2 statistic with continuity correction is $T_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = 142.3621$.

number of nodes		network 1 present	network 1 absent
network 2	present	15	0
	absent	5	445

Table 29: Theoretical contingency table for simulated networks with 20 additional nodes. Value of the χ^2 statistic with continuity correction is $T_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = 321.2684$.

number of nodes		network 1 present	network 1 absent
network 2	present	15	0
	absent	5	1810

Table 30: Theoretical contingency table for simulated networks with 50 additional nodes. Value of the χ^2 statistic with continuity correction is $T_5\left(\vec{\rho}_1, \vec{\rho}_2\right) = 1278.017$.

		Cor − pCor with shrinkage					pCor with − pCor without shrinkage				
	variance	number of additional nodes					number of additional nodes				
		0	5	10	20	50	0	5	10	20	50
MaxDA	0.01	0.17	0.27	0.38	0.53	0.52	0.10	0.04	-0.01	-0.11	0.07
	0.1	0.05	0.22	0.23	0.31	0.34	0.18	0.10	0.15	0.03	0.16
	0.5	0.00	0.09	0.24	0.35	0.52	0.21	0.11	0.07	0.01	0.08
	1	0.02	0.06	0.26	0.46	0.55	0.16	0.08	0.14	-0.09	0.04
	2	-0.01	0.11	0.24	0.48	0.51	0.29	0.30	0.16	0.06	0.09
	4	0.06	0.12	0.39	0.48	0.57	0.37	0.28	0.19	0.04	0.11
	8	0.15	0.25	0.32	0.43	0.26	0.24	0.30	0.17	0.17	0.40
	16	0.17	0.25	0.30	0.27	0.23	0.35	0.25	0.31	0.22	0.32
MDA	0.01	0.10	0.12	0.10	0.15	0.14	0.47	0.54	0.61	0.65	0.64
	0.1	0.06	0.06	0.03	0.04	0.06	0.46	0.59	0.64	0.65	0.60
	0.5	0.01	0.05	0.02	0.09	0.01	0.49	0.58	0.66	0.59	0.68
	1	0.02	0.10	0.11	0.15	0.02	0.48	0.45	0.56	0.60	0.53
	2	0.18	0.21	0.17	0.31	-0.01	0.38	0.48	0.53	0.36	0.55
	4	0.30	0.34	0.50	0.34	-0.10	0.37	0.23	0.23	0.35	0.61
	8	0.48	0.39	0.45	0.19	-0.12	0.10	0.24	0.19	0.28	0.37
	16	0.43	0.37	0.26	0.08	-0.04	0.17	0.24	0.25	0.21	0.11
MDQ	0.01	0.14	0.15	0.15	0.27	0.17	0.27	0.29	0.47	0.50	0.70
	0.1	0.05	0.07	0.04	0.13	0.12	0.27	0.42	0.55	0.57	0.60
	0.5	0.02	0.02	0.01	0.13	0.10	0.38	0.48	0.59	0.50	0.68
	1	0.01	0.04	0.04	0.18	0.13	0.36	0.45	0.51	0.53	0.57
	2	0.08	0.06	0.13	0.27	0.15	0.35	0.54	0.55	0.41	0.55
	4	0.06	0.15	0.37	0.32	0.13	0.51	0.40	0.35	0.42	0.56
	8	0.25	0.29	0.39	0.26	0.18	0.27	0.31	0.28	0.32	0.39
	16	0.30	0.32	0.32	0.22	0.10	0.29	0.39	0.30	0.24	0.14

Table 31: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 200 samples per group.

		Cor − pCor with shrinkage					pCor with − pCor without shrinkage				
	variance	number of additional nodes					number of additional nodes				
		0	5	10	20	50	0	5	10	20	50
MaxDA	001	0.25	0.45	0.58	0.57	0.34	0.06	-0.21	-0.27	-0.31	0.03
	01	0.03	0.37	0.40	0.42	0.29	0.21	-0.12	-0.11	-0.23	-0.06
	05	0.07	0.29	0.45	0.28	0.30	0.11	-0.11	-0.15	0.03	0.04
	1	0.03	0.21	0.30	0.35	0.32	0.08	-0.06	-0.08	-0.18	0.19
	2	0.06	0.17	0.33	0.47	0.26	0.06	0.00	-0.12	-0.11	0.21
	4	0.05	0.14	0.36	0.39	0.24	0.14	0.09	-0.02	-0.04	0.16
	8	0.14	0.28	0.33	0.29	0.36	0.19	0.03	0.14	0.18	0.25
	16	0.22	0.26	0.23	0.19	0.04	0.16	0.11	0.13	0.26	0.29
MDA	001	0.25	0.19	0.21	0.57	0.58	0.36	0.41	0.48	0.16	-0.05
	01	0.04	0.02	0.14	0.59	0.55	0.43	0.49	0.54	0.08	0.01
	05	0.08	0.19	0.38	0.71	0.63	0.32	0.44	0.21	-0.05	-0.10
	1	0.23	0.30	0.52	0.67	0.57	0.30	0.21	0.03	-0.07	-0.07
	2	0.31	0.52	0.63	0.64	0.45	0.15	0.00	0.01	-0.10	-0.03
	4	0.49	0.59	0.69	0.53	0.36	0.11	-0.02	-0.16	-0.01	-0.05
	8	0.51	0.56	0.58	0.44	0.23	0.08	0.06	0.04	-0.04	0.00
	16	0.51	0.37	0.25	0.23	0.05	-0.02	0.00	0.00	-0.03	0.11
MDQ	001	0.25	0.21	0.22	0.58	0.79	0.22	0.24	0.24	0.02	-0.11
	01	0.04	0.01	0.11	0.60	0.71	0.30	0.31	0.39	-0.10	-0.07
	05	0.03	0.08	0.27	0.64	0.75	0.19	0.38	0.22	-0.10	-0.09
	1	0.04	0.14	0.40	0.64	0.75	0.32	0.22	0.05	-0.18	-0.07
	2	0.13	0.22	0.48	0.72	0.65	0.18	0.08	0.05	-0.16	0.00
	4	0.29	0.33	0.68	0.64	0.62	0.15	0.08	-0.16	0.01	-0.05
	8	0.37	0.52	0.64	0.57	0.44	0.10	0.11	0.03	-0.01	0.01
	16	0.40	0.43	0.33	0.36	0.15	0.10	0.04	0.03	0.03	0.07

Table 32: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 300 samples in the first and 100 samples in the second group.

		Cor − pCor with shrinkage				pCor with − pCor without shrinkage			
	variance	number of additional nodes				number of additional nodes			
		0	5	10	20	0	5	10	20
MaxDA	0.01	0.24	0.60	0.71	0.52	0.47	0.13	-0.31	-0.13
	0.1	0.17	0.57	0.76	0.57	0.50	0.15	-0.24	-0.08
	0.5	0.26	0.55	0.68	0.63	0.44	0.19	-0.26	0.03
	1	0.26	0.46	0.69	0.61	0.51	0.19	-0.04	0.00
	2	0.28	0.53	0.63	0.56	0.44	0.17	0.08	0.02
	4	0.25	0.57	0.53	0.52	0.45	0.04	0.05	0.06
	8	0.49	0.33	0.40	0.42	0.10	0.19	0.06	0.03
	16	0.27	0.27	0.16	0.14	0.14	0.17	0.09	0.15
MDA	0.01	0.17	0.08	0.24	0.62	0.61	0.69	0.45	0.01
	0.1	0.04	0.04	0.26	0.59	0.69	0.73	0.44	0.00
	0.5	0.15	0.33	0.47	0.53	0.59	0.41	0.14	-0.07
	1	0.33	0.59	0.55	0.58	0.44	0.08	0.08	-0.02
	2	0.53	0.54	0.50	0.46	0.21	0.13	0.02	-0.04
	4	0.57	0.47	0.42	0.35	0.12	0.06	0.02	0.03
	8	0.41	0.29	0.24	0.31	0.12	0.13	0.01	0.03
	16	0.26	0.28	0.17	0.06	0.11	0.01	0.00	0.07
MDQ	0.01	0.18	0.17	0.49	0.77	0.63	0.62	0.24	0.01
	0.1	0.04	0.12	0.59	0.67	0.69	0.68	0.15	-0.03
	0.5	0.08	0.29	0.54	0.69	0.65	0.47	0.08	-0.06
	1	0.17	0.51	0.66	0.72	0.61	0.18	0.07	-0.04
	2	0.37	0.55	0.61	0.64	0.37	0.15	0.04	-0.04
	4	0.48	0.55	0.52	0.51	0.24	0.06	0.05	0.01
	8	0.50	0.41	0.44	0.37	0.13	0.15	0.02	0.05
	16	0.27	0.35	0.18	0.16	0.15	0.05	0.03	0.08

Table 33: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 150 samples in the first and 50 samples in the second group.

		Cor − pCor with shrinkage				pCor with − pCor without shrinkage			
	variance	number of additional nodes				number of additional nodes			
		0	5	10	20	0	5	10	20
MaxDA	0.01	0.23	0.64	0.56	0.43	0.43	-0.04	-0.07	-0.04
	0.1	0.12	0.59	0.57	0.53	0.56	0.04	-0.17	-0.07
	0.5	0.16	0.61	0.67	0.53	0.57	-0.04	-0.13	0.02
	1	0.21	0.54	0.68	0.55	0.40	0.06	-0.05	0.00
	2	0.30	0.53	0.57	0.37	0.32	0.05	-0.05	0.06
	4	0.36	0.53	0.41	0.33	0.27	0.00	-0.02	0.03
	8	0.36	0.41	0.29	0.34	0.08	0.03	0.07	-0.01
	16	0.25	0.12	0.09	0.09	0.04	0.05	0.02	0.01
MDA	0.01	0.08	-0.05	0.02	0.11	0.65	0.73	0.66	0.36
	0.1	0.01	-0.03	0.00	0.07	0.68	0.70	0.52	0.40
	0.5	-0.04	-0.04	0.06	0.04	0.71	0.58	0.48	0.35
	1	-0.01	-0.01	0.12	-0.03	0.61	0.56	0.43	0.37
	2	0.18	0.11	0.09	0.03	0.45	0.37	0.32	0.34
	4	0.29	0.23	0.11	-0.03	0.28	0.14	0.20	0.28
	8	0.17	0.11	0.02	-0.06	0.18	0.21	0.24	0.27
	16	0.02	-0.01	0.01	-0.02	0.12	0.08	0.10	0.07
MDQ	0.01	0.10	0.27	0.33	0.36	0.66	0.48	0.43	0.22
	0.1	0.02	0.14	0.36	0.30	0.70	0.57	0.24	0.30
	0.5	-0.01	0.17	0.33	0.28	0.71	0.44	0.29	0.27
	1	0.00	0.25	0.48	0.23	0.65	0.36	0.24	0.35
	2	0.19	0.30	0.32	0.12	0.47	0.32	0.24	0.35
	4	0.25	0.32	0.27	0.16	0.35	0.21	0.15	0.25
	8	0.27	0.18	0.18	0.05	0.19	0.25	0.23	0.22
	16	0.07	0.07	0.06	0.01	0.07	0.10	0.08	0.08

Table 34: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 50 samples per group.

		Cor − pCor with shrinkage		pCor with − pCor without shrinkage	
	variance	number of additional nodes		number of additional nodes	
		0	5	0	5
MaxDA	0.01	0.51	0.60	0.04	-0.01
	0.1	0.37	0.56	0.19	0.08
	0.5	0.50	0.46	0.14	0.07
	1	0.41	0.41	0.23	0.14
	2	0.35	0.37	0.17	0.11
	4	0.42	0.34	0.06	-0.01
	8	0.37	0.21	-0.04	0.06
	16	0.17	0.17	0.05	0.01
MDA	0.01	0.50	0.60	0.15	0.11
	0.1	0.51	0.64	0.09	0.04
	0.5	0.64	0.55	0.05	0.03
	1	0.60	0.53	0.07	0.01
	2	0.43	0.43	0.06	0.03
	4	0.34	0.38	0.02	-0.05
	8	0.34	0.18	-0.01	-0.01
	16	0.10	0.05	0.07	0.03
MDQ	0.01	0.45	0.51	0.23	0.22
	0.1	0.45	0.59	0.20	0.11
	0.5	0.60	0.57	0.10	0.06
	1	0.57	0.56	0.12	0.01
	2	0.43	0.56	0.13	-0.02
	4	0.37	0.39	0.07	-0.01
	8	0.47	0.23	0.00	0.00
	16	0.13	0.08	0.07	0.03

Table 35: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 180 samples in the first and 20 samples in the second group.

		Cor − pCor with shrinkage		pCor with − pCor without shrinkage	
	variance	number of additional nodes		number of additional nodes	
		0	5	0	5
MaxDA	0.01	0.20	0.27	0.12	0.01
	0.1	0.31	0.29	0.09	0.02
	0.5	0.19	0.19	0.25	0.00
	1	0.20	0.12	0.10	0.05
	2	0.22	0.22	0.05	-0.02
	4	0.20	0.10	0.08	0.01
	8	0.10	0.08	-0.01	0.04
	16	0.04	0.02	-0.03	0.02
MDA	0.01	-0.24	-0.02	0.62	0.46
	0.1	-0.36	-0.13	0.69	0.48
	0.5	-0.24	-0.06	0.59	0.34
	1	-0.09	-0.02	0.45	0.27
	2	-0.01	0.03	0.29	0.18
	4	0.04	-0.02	0.13	0.07
	8	0.06	-0.03	0.01	0.08
	16	0.01	0.02	-0.02	0.02
MDQ	0.01	-0.08	0.35	0.57	0.19
	0.1	-0.19	0.29	0.65	0.16
	0.5	-0.18	0.27	0.57	0.11
	1	0.03	0.18	0.42	0.12
	2	0.15	0.16	0.26	0.16
	4	0.08	0.08	0.15	0.07
	8	0.10	0.01	0.01	0.06
	16	0.01	0.02	-0.05	0.03

Table 36: Comparison of power for the tests using MaxDA·, MDA·, and MDQ· as test statistics based on partial correlations (pCor) with and without shrinkage of the covariance matrix and ordinary correlations (Cor) considering 20 samples per group.

variance	number of additional nodes				
	0	5	10	20	50
	max MDQ - max MDA				
0.01	0.01	0.03	0.04	0.06	0.09
0.1	0.01	0.01	0.04	0.03	0.14
0.5	0.00	0.00	0.01	0.10	0.12
1	0.03	0.00	0.06	0.10	0.05
2	0.00	0.02	0.03	0.11	0.16
4	0.05	0.09	0.13	0.15	0.09
8	0.03	0.10	0.09	0.21	0.05
16	0.09	0.07	0.14	0.12	0.03
	max MDA - max MaxDA				
0.01	-0.02	0.03	-0.03	-0.01	-0.06
0.1	0.02	0.00	-0.01	-0.02	-0.12
0.5	0.01	0.00	-0.01	-0.11	-0.20
1	-0.02	-0.02	-0.06	-0.09	-0.13
2	-0.03	-0.01	-0.04	-0.16	-0.23
4	-0.07	-0.06	-0.09	-0.20	-0.19
8	-0.07	-0.10	-0.15	-0.24	-0.25
16	-0.15	-0.21	-0.24	-0.26	-0.23
	max MaxDA - max MDQ				
0.01	0.01	-0.06	-0.01	-0.05	-0.03
0.1	-0.03	-0.01	-0.03	-0.01	-0.02
0.5	-0.01	0.00	0.00	0.01	0.08
1	-0.01	0.02	0.00	-0.01	0.08
2	0.03	-0.01	0.01	0.05	0.07
4	0.02	-0.03	-0.04	0.05	0.10
8	0.04	0.00	0.06	0.03	0.20
16	0.06	0.14	0.10	0.14	0.20

Table 37: Comparison of maximal power using on partial correlations (pCor) with or without shrinkage of the covariance matrix or ordinary correlations (Cor) of the tests using MaxDA·, MDA·, and MDQ· as test statistics considering 100 samples per group.

variance	number of additional nodes	
	0	5
	max MDQ - max MDA	
001	-0.06	0.08
01	-0.05	-0.01
05	-0.02	0.03
1	-0.01	0.04
2	0.13	0.10
4	0.08	0.06
8	0.04	-0.01
16	0.03	0.00
	max MDA - max MaxDA	
001	0.24	0.19
01	0.26	0.19
05	0.20	0.13
1	0.12	0.13
2	-0.02	0.00
4	-0.11	-0.02
8	-0.03	-0.02
16	-0.01	0.02
	max MaxDA - max MDQ	
001	-0.18	-0.27
01	-0.21	-0.18
05	-0.18	-0.16
1	-0.11	-0.17
2	-0.11	-0.10
4	0.03	-0.04
8	-0.01	0.03
16	-0.02	-0.02

Table 38: Comparison of maximal power using on partial correlations (pCor) with or without shrinkage of the covariance matrix or ordinary correlations (Cor) of the tests using MaxDA·, MDA·, and MDQ· as test statistics considering 20 samples per group.

Table 39: Overview of the results of the permutation tests for the enriched GO groups in the ontologies biological process (top), molecular function (middle), and cellular component (bottom), separated by midrules, regarding MammaPrint genes. p-values below 0.05 are marked in grey.

GO term	MaxDApC	MDApC	MDQpC	MDE	χE	MDAR	MDQR	MDARE	MDQRE	CORpCE	RCORpCE	MaxDAC	MDAC	MDQC
GO:0030949	0.936	0.045	0.071	0.978	0.966	0.050	0.057	0.692	0.661	0.547	0.273	0.210	0.022	0.024
GO:0044342	0.507	0.194	0.347	0.835	0.615	0.128	0.053	0.858	0.858	0.874	0.703	0.059	0.287	0.147
GO:0048630	0.098	0.845	0.857	0.036	0.810	0.797	0.671	1.000	1.000	0.392	0.977	0.312	0.672	0.632
GO:0001957	0.082	0.099	0.067	0.955	0.694	0.552	0.583	0.504	0.467	0.027	0.045	0.287	0.015	0.028
GO:0043569	0.288	0.918	0.902	0.815	0.031	0.588	0.657	0.888	0.873	0.020	0.072	0.696	0.710	0.682
GO:0014912	0.112	0.116	0.067	0.825	0.534	0.509	0.565	0.865	0.759	0.059	0.615	0.566	0.085	0.116
GO:0002063	0.541	0.266	0.303	0.367	0.381	0.178	0.158	0.383	0.336	0.315	0.474	0.408	0.060	0.101
GO:0032332	0.624	0.003	0.004	0.955	0.750	0.005	0.003	0.660	0.579	0.001	0.010	0.252	0.009	0.005
GO:0060056	0.446	0.698	0.814	0.036	0.006	0.228	0.379	0.704	0.710	0.968	0.908	0.161	0.522	0.560
GO:0040001	0.068	0.305	0.291	0.929	0.756	0.442	0.245	0.936	0.826	0.077	0.112	0.075	0.031	0.030
GO:0043568	0.112	0.508	0.775	0.873	0.724	0.032	0.013	0.854	0.840	0.359	0.665	0.244	0.351	0.262
GO:0071320	0.473	0.331	0.486	0.073	0.112	0.150	0.172	0.311	0.357	0.755	0.896	0.744	0.542	0.562
GO:0045668	0.715	0.244	0.162	0.049	0.022	0.245	0.137	0.034	0.027	0.132	0.090	0.528	0.308	0.279
GO:0031069	0.361	0.221	0.049	0.591	0.261	0.372	0.312	0.616	0.598	0.130	0.268	0.785	0.758	0.778
GO:0001958	0.571	0.528	0.628	0.544	0.586	0.404	0.423	0.739	0.702	0.947	0.841	0.377	0.343	0.345
GO:0032508	0.865	0.915	0.970	0.909	0.936	0.518	0.527	0.841	0.854	0.495	0.870	0.630	0.328	0.367
GO:0071407	0.352	0.070	0.099	0.768	0.802	0.477	0.510	0.625	0.673	0.186	0.613	0.238	0.218	0.221
GO:0017148	0.841	0.320	0.424	0.073	0.269	0.475	0.472	0.146	0.117	0.493	0.705	0.045	0.004	0.007
GO:0010508	0.162	0.649	0.331	0.984	0.799	0.948	0.946	0.956	0.956	0.564	0.481	0.278	0.230	0.186
GO:0038084	0.926	0.475	0.647	0.326	0.615	0.127	0.104	0.566	0.566	0.373	0.611	0.087	0.036	0.031
GO:0019852	0.486	0.579	0.717	0.306	0.619	0.410	0.312	0.200	0.213	0.597	0.746	0.032	0.450	0.099
GO:0070830	0.466	0.467	0.565	0.845	0.929	0.082	0.088	0.536	0.560	0.753	0.772	0.197	0.299	0.295
GO:0006940	0.708	0.579	0.658	0.636	0.693	0.464	0.351	0.696	0.656	0.671	0.674	0.864	0.368	0.431
GO:0031532	0.790	0.016	0.023	0.058	0.068	0.055	0.027	0.019	0.023	0.118	0.095	0.227	0.135	0.167
GO:0031032	0.314	0.051	0.056	0.941	0.876	0.718	0.565	0.866	0.891	0.096	0.373	0.325	0.340	0.369
GO:0030901	0.792	0.373	0.259	0.953	0.801	0.878	0.899	0.938	0.918	0.842	0.426	0.466	0.409	0.385
GO:0043552	0.512	0.132	0.185	0.780	0.689	0.088	0.110	0.546	0.508	0.088	0.140	0.516	0.082	0.081
GO:0002548	0.487	0.022	0.045	0.265	0.226	0.003	0.002	0.253	0.264	0.403	0.403	0.888	0.205	0.272
GO:0021670	0.839	0.657	0.766	0.070	0.163	0.042	0.052	0.068	0.070	0.634	0.503	0.546	0.129	0.112
GO:0051382	0.512	0.516	0.622	0.462	0.410	0.437	0.336	0.272	0.244	0.252	0.279	0.671	0.685	0.603
GO:0000281	0.101	0.111	0.148	0.724	0.212	0.055	0.030	0.676	0.621	0.139	0.168	0.420	0.146	0.196
GO:0031663	0.952	0.059	0.088	0.021	0.026	0.336	0.423	0.039	0.043	0.482	0.555	0.199	0.024	0.025

Table 39: Overview of the results of the permutation tests for the enriched GO groups in the ontologies biological process (top), molecular function (middle), and cellular component (bottom), separated by midrules, regarding MammaPrint genes. p-values below 0.05 are marked in grey.

GO term	MaxDApC	MDApC	MDQpC	MDE	χE	MDAR	MDQR	MDARE	MDQRE	CORpCE	RCORpCE	MaxDAC	MDAC	MDQC
GO:0007076	0.295	0.182	0.085	0.973	0.202	0.637	0.778	0.904	0.904	0.003	0.005	0.131	0.060	0.064
GO:0048853	0.116	0.071	0.070	0.656	0.210	0.096	0.097	0.914	0.914	0.186	0.616	0.105	0.016	0.013
GO:0008608	0.178	0.026	0.047	0.710	0.778	0.108	0.219	0.832	0.836	0.824	0.935	0.003	0.028	0.030
GO:0000186	0.328	0.010	0.005	0.103	0.174	0.077	0.061	0.192	0.153	0.464	0.777	0.394	0.010	0.010
GO:0014068	0.109	0.050	0.044	0.411	0.383	0.246	0.086	0.372	0.359	0.212	0.331	0.331	0.020	0.023
GO:0034080	0.726	0.354	0.697	0.408	0.405	0.004	0.009	0.145	0.142	0.205	0.098	0.894	0.492	0.531
GO:0021772	0.740	0.966	0.981	0.875	0.819	0.381	0.437	0.770	0.737	0.481	0.560	0.885	0.313	0.362
GO:0006024	0.869	0.115	0.065	0.617	0.557	0.381	0.404	0.557	0.391	0.195	0.222	0.319	0.030	0.022
GO:0005520	0.511	0.179	0.191	0.396	0.493	0.166	0.075	0.225	0.217	0.253	0.095	0.193	0.104	0.088
GO:0001968	0.085	0.016	0.031	0.207	0.101	0.018	0.012	0.220	0.164	0.584	0.158	0.362	0.258	0.285
GO:0003678	0.866	0.788	0.877	0.857	0.831	0.057	0.034	0.841	0.853	0.385	0.757	0.266	0.095	0.111
GO:0005355	0.185	0.726	0.777	0.295	0.654	0.362	0.309	0.113	0.156	0.679	0.792	0.063	0.755	0.607
GO:0005021	0.934	0.452	0.640	0.353	0.624	0.114	0.081	0.607	0.607	0.371	0.608	0.101	0.028	0.027
GO:0051059	0.289	0.183	0.253	0.960	0.508	0.071	0.291	0.998	0.998	0.463	0.901	0.390	0.331	0.305
GO:0070628	0.191	0.001	0.001	0.364	0.346	0.017	0.009	0.540	0.433	0.160	0.651	0.029	0.005	0.010
GO:0004029	0.473	0.318	0.205	1.000	0.006	0.326	0.210	1.000	1.000	0.995	0.995	0.257	0.235	0.213
GO:0030276	0.599	0.214	0.220	0.910	0.840	0.433	0.302	0.723	0.803	0.242	0.388	0.551	0.469	0.448
GO:0008242	0.780	0.806	0.647	0.400	0.776	0.886	0.704	0.311	0.241	0.027	0.003	0.541	0.461	0.334
GO:0001085	0.087	0.022	0.019	0.119	0.264	0.013	0.052	0.078	0.090	0.236	0.311	0.447	0.115	0.087
GO:0000145	0.843	0.907	0.917	0.783	0.499	0.972	0.954	0.478	0.463	0.607	0.140	0.872	0.814	0.736
GO:0016942	0.188	0.074	0.147	0.492	0.303	0.060	0.081	0.914	0.835	0.908	0.751	0.228	0.168	0.158
GO:0034451	0.985	0.935	0.979	0.992	0.885	0.394	0.439	0.953	0.942	0.731	0.055	0.997	0.852	0.912
GO:0046581	0.135	0.339	0.302	0.676	0.092	0.690	0.841	0.524	0.511	0.008	0.012	0.339	0.615	0.569
GO:0005923	0.402	0.122	0.141	0.155	0.545	0.219	0.226	0.166	0.237	0.621	0.874	0.710	0.325	0.322
GO:0005587	0.294	0.439	0.220	0.775	0.703	0.862	0.802	0.134	0.098	0.548	0.058	0.250	0.027	0.033
GO:0031011	0.131	0.184	0.202	0.977	0.876	0.684	0.678	0.617	0.528	0.171	0.095	0.387	0.031	0.038

Additional figures

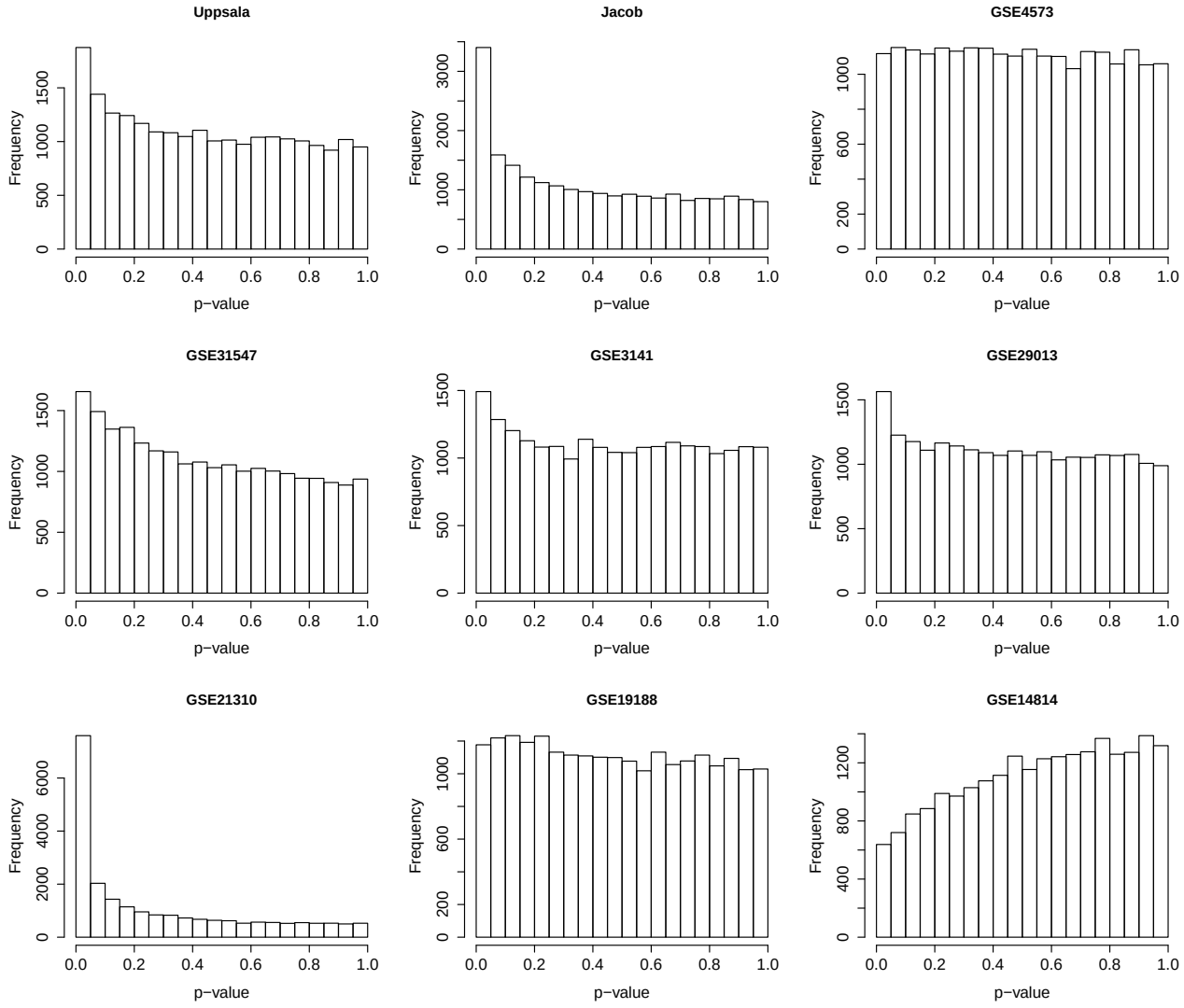

Figure 19: Histogramms of the p-values of the Wald tests testing the hypothesis "H_0 : HR $= 1$" for every probe set for all nine non-small cell lung cancer datasets.

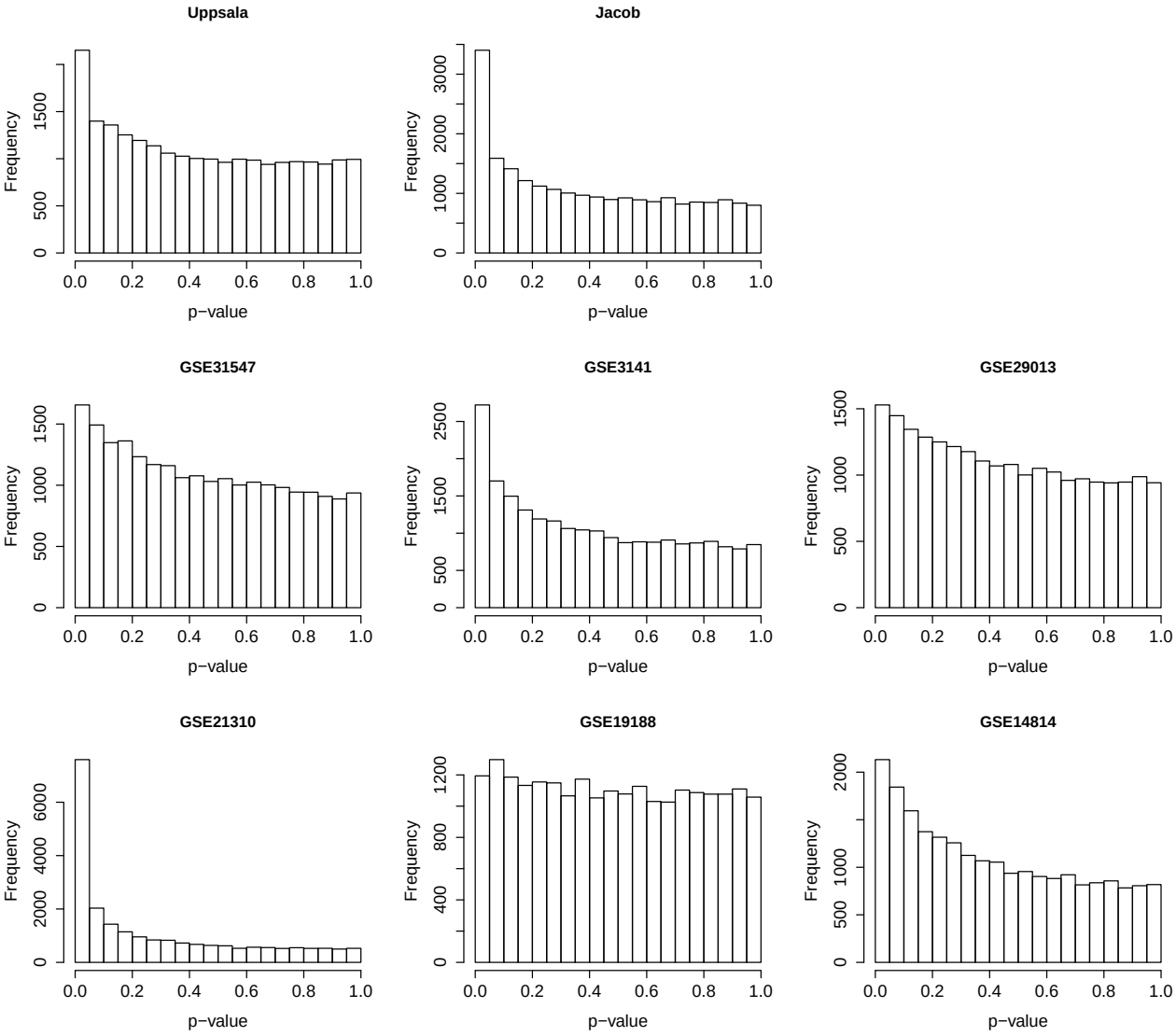

Figure 20: Histogramms of the p-values of the Wald tests testing the hypothesis "H_0 : HR $= 1$" for every probe set in the histological subgroup of adenocarcinomas for the non-small cell lung cancer datasets.

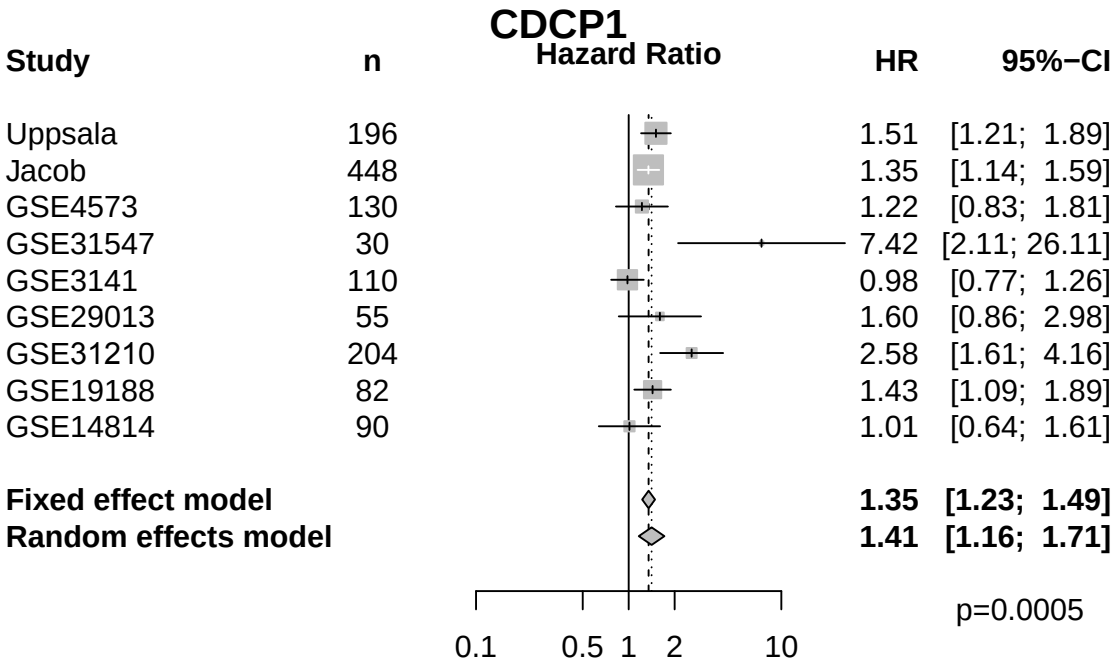

Figure 21: Forestplot of the meta-analysis for probe set "218451_at" that represents the gene "CDCP1" including all nine non-small cell lung cancer datasets. The p-value (bottom right) corresponds to the random effects model.

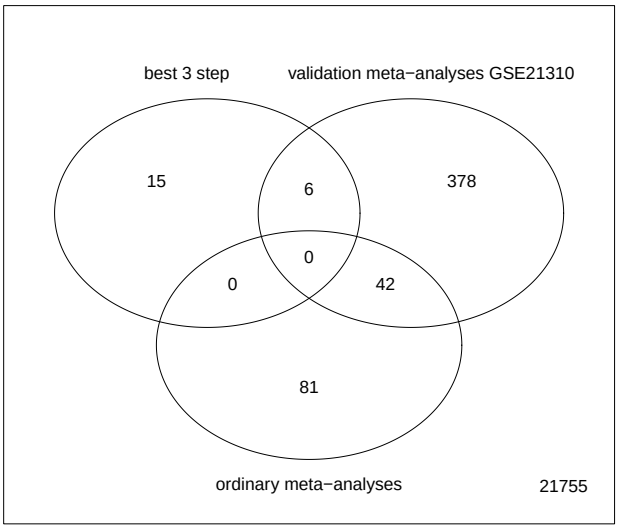

Figure 22: Visualisation of the significant features at the end of the best 3-step sequential validation approach, the combined meta-analysis for all patients comparing the proceeding when GSE31210 is used as basic dataset, and the common meta-analyses assuming a random effects model.

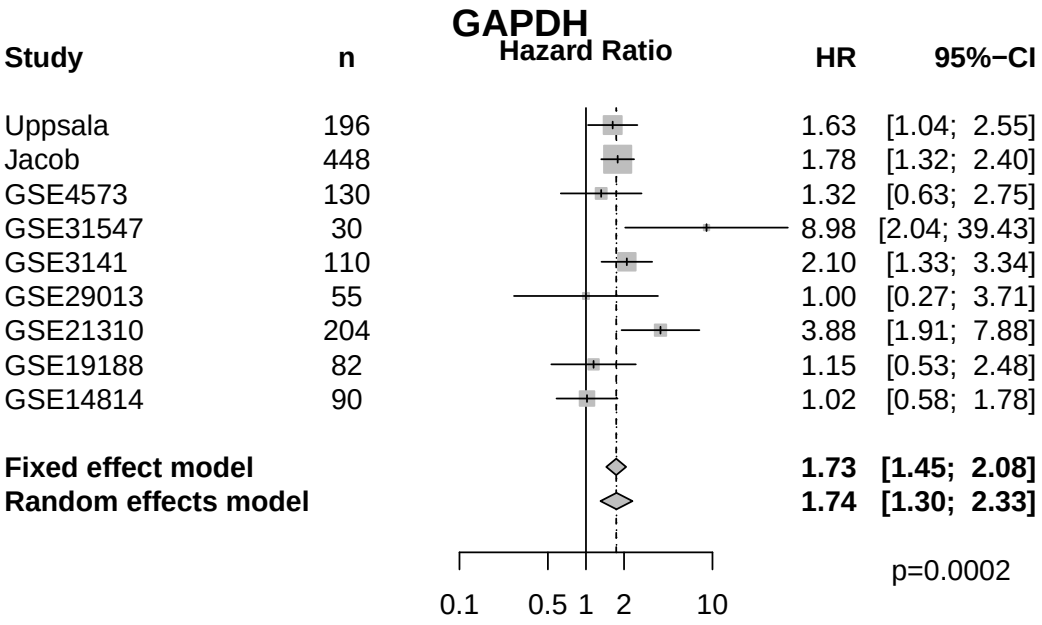

Figure 23: Forestplot of the meta-analysis for probe set "212581_x_at" that represents the gene "GAPDH" including all nine non-small cell lung cancer datasets. The p-value (bottom right) corresponds to the random effects model.

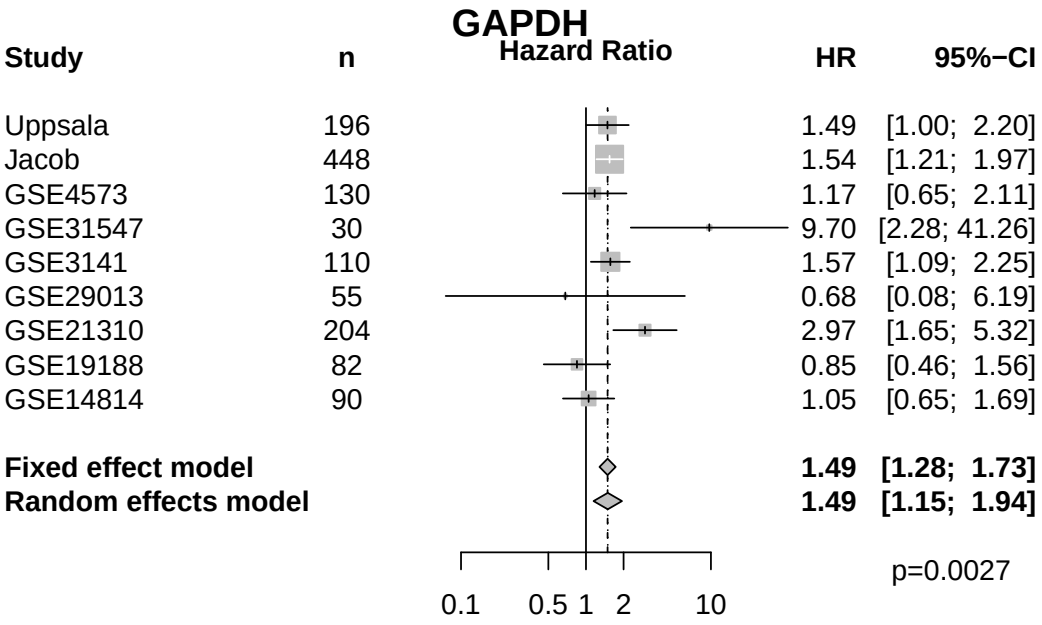

Figure 24: Forestplot of the meta-analysis for probe set "M33197_M_at" that represents the gene "GAPDH" including all nine non-small cell lung cancer datasets. The p-value (bottom right) corresponds to the random effects model.

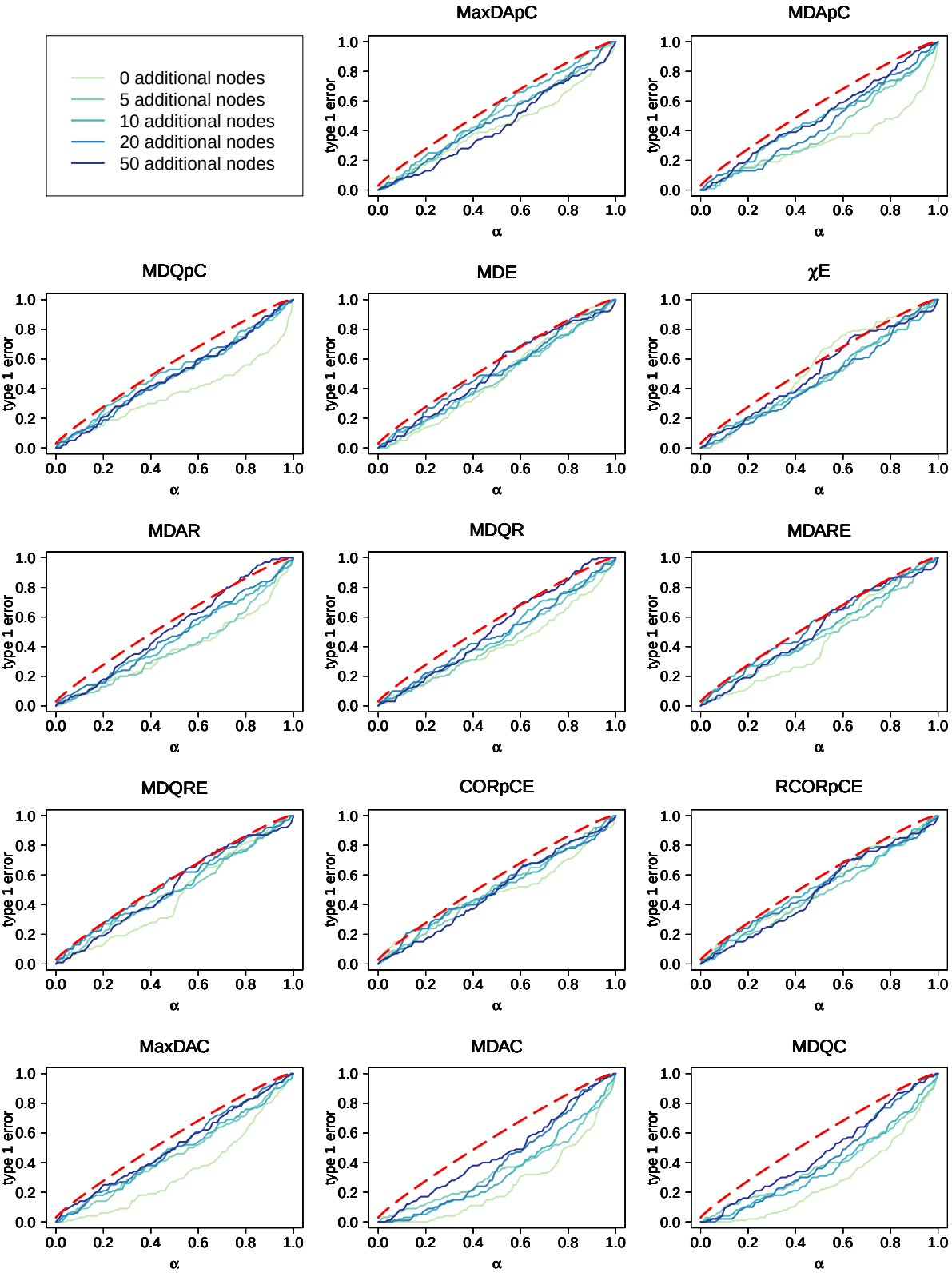

Figure 25: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 0.01.

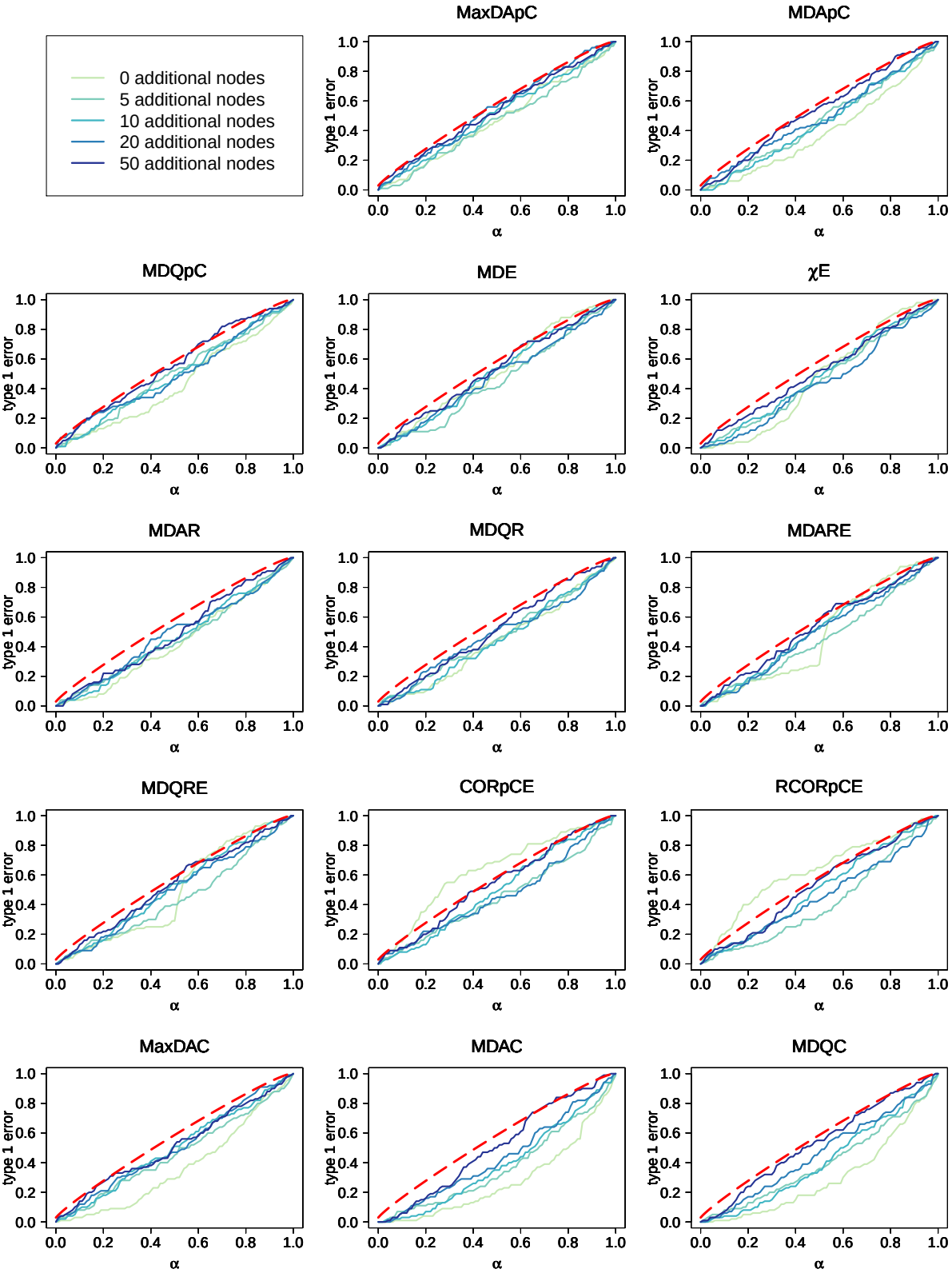

Figure 26: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 0.1.

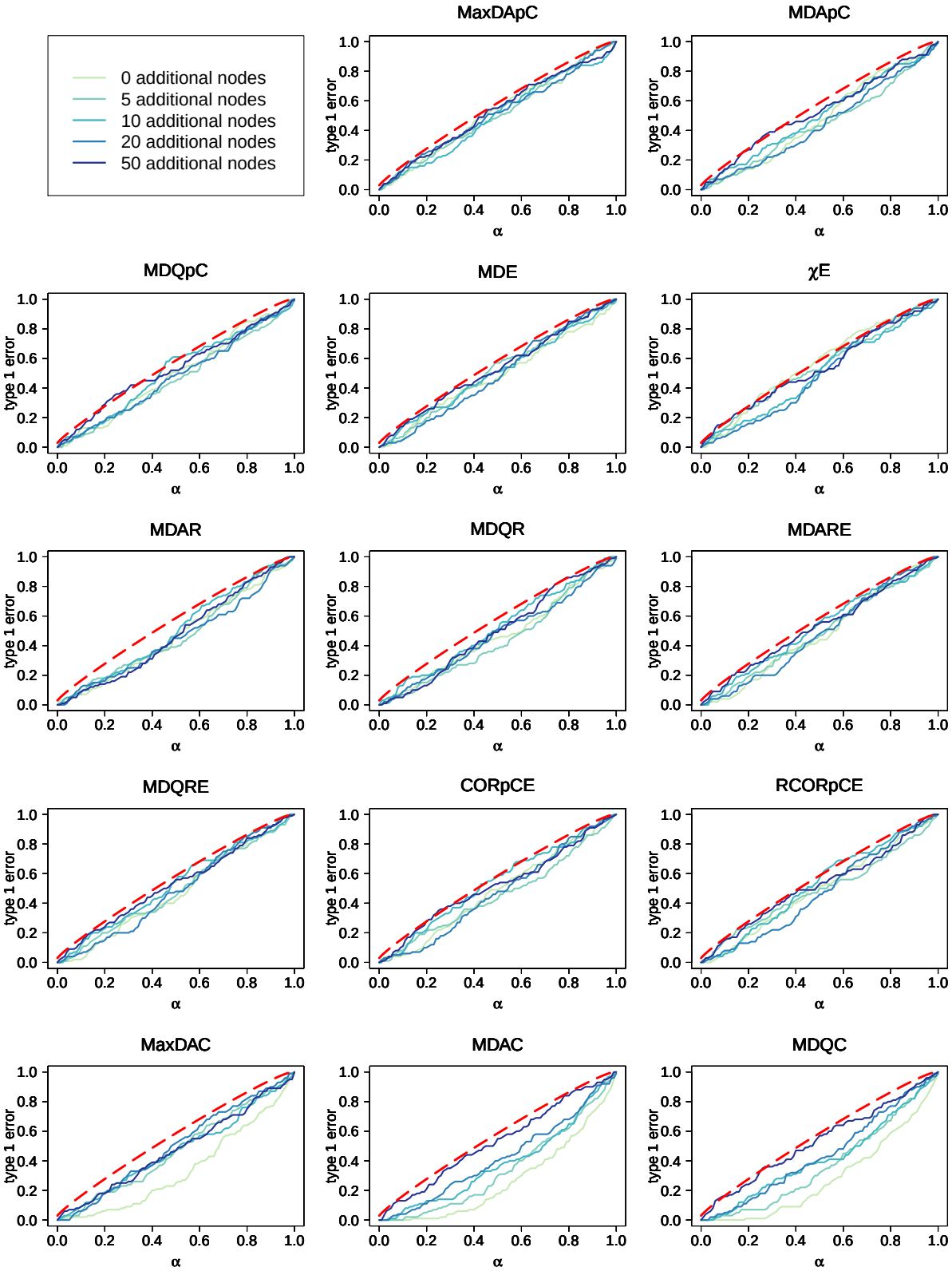

Figure 27: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 0.5.

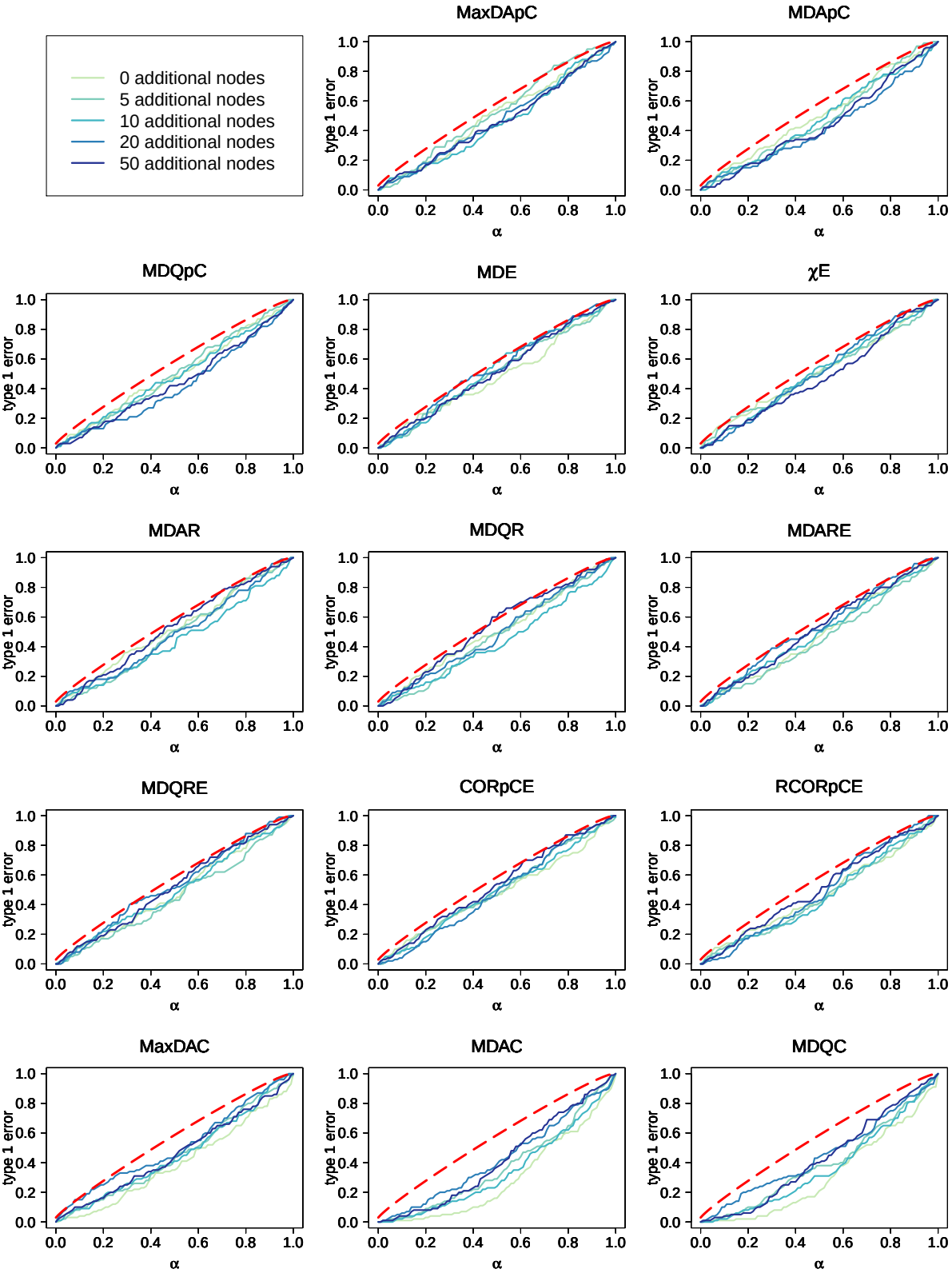

Figure 28: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 1.

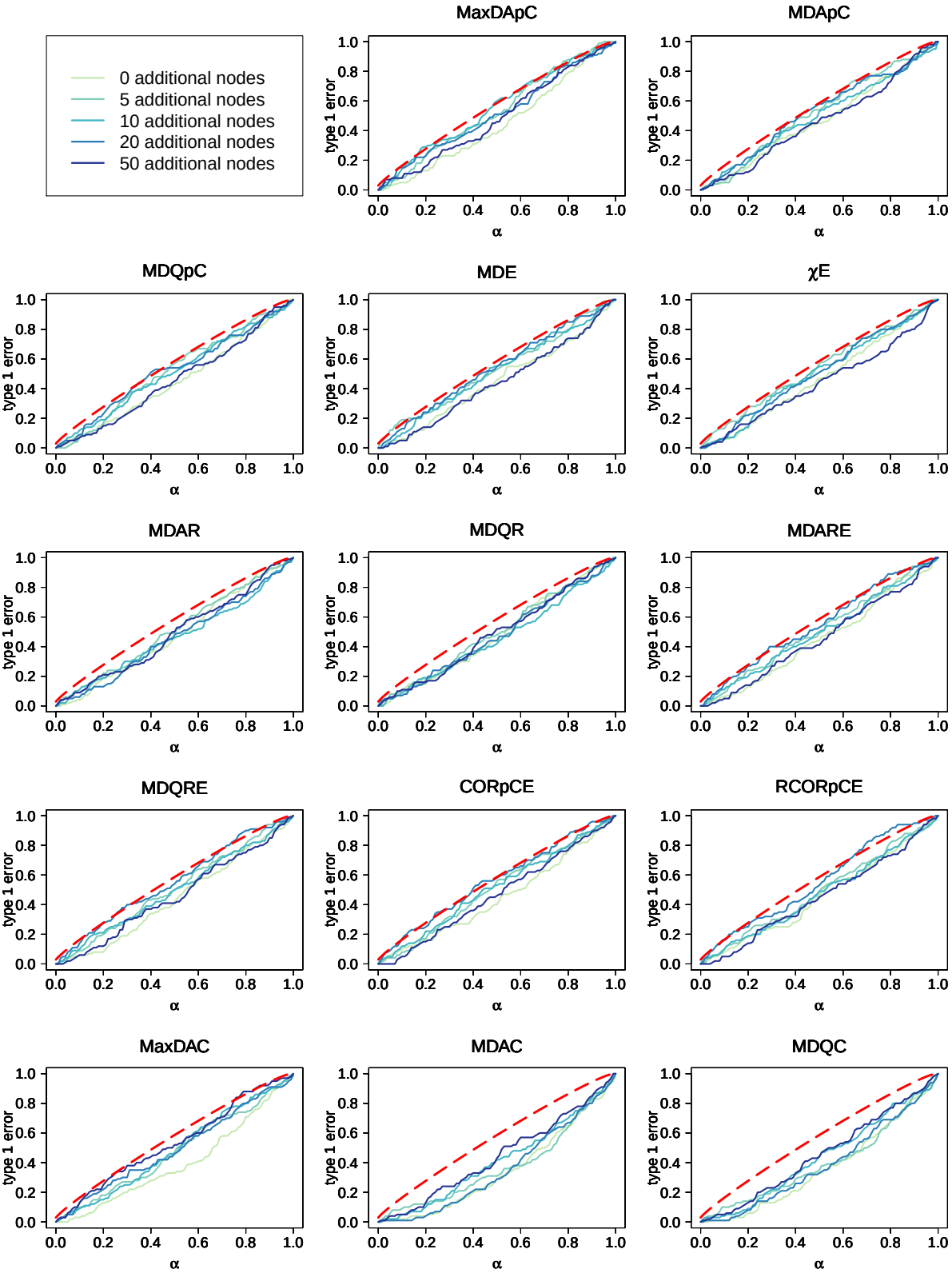

Figure 29: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 2.

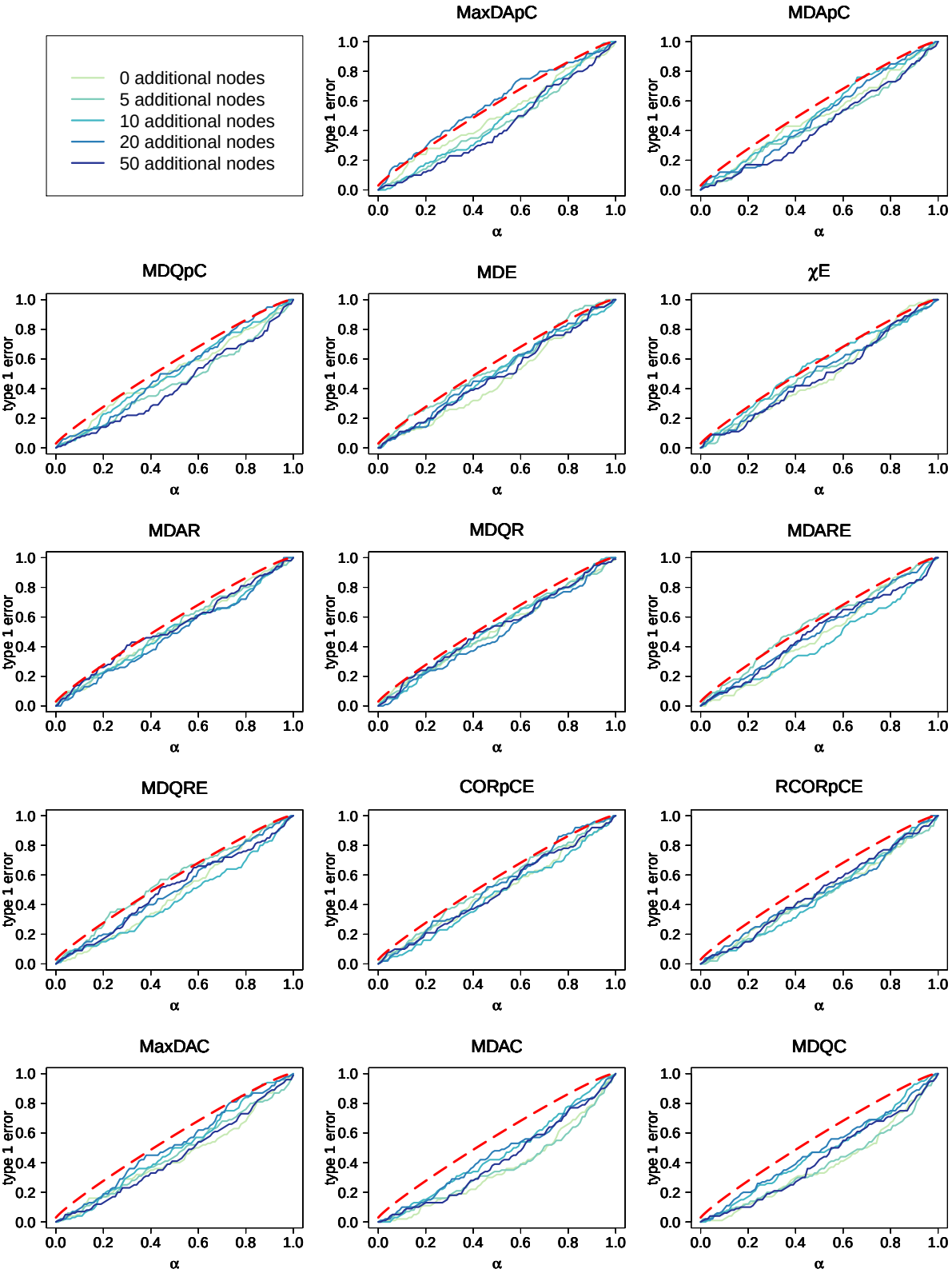

Figure 30: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 4.

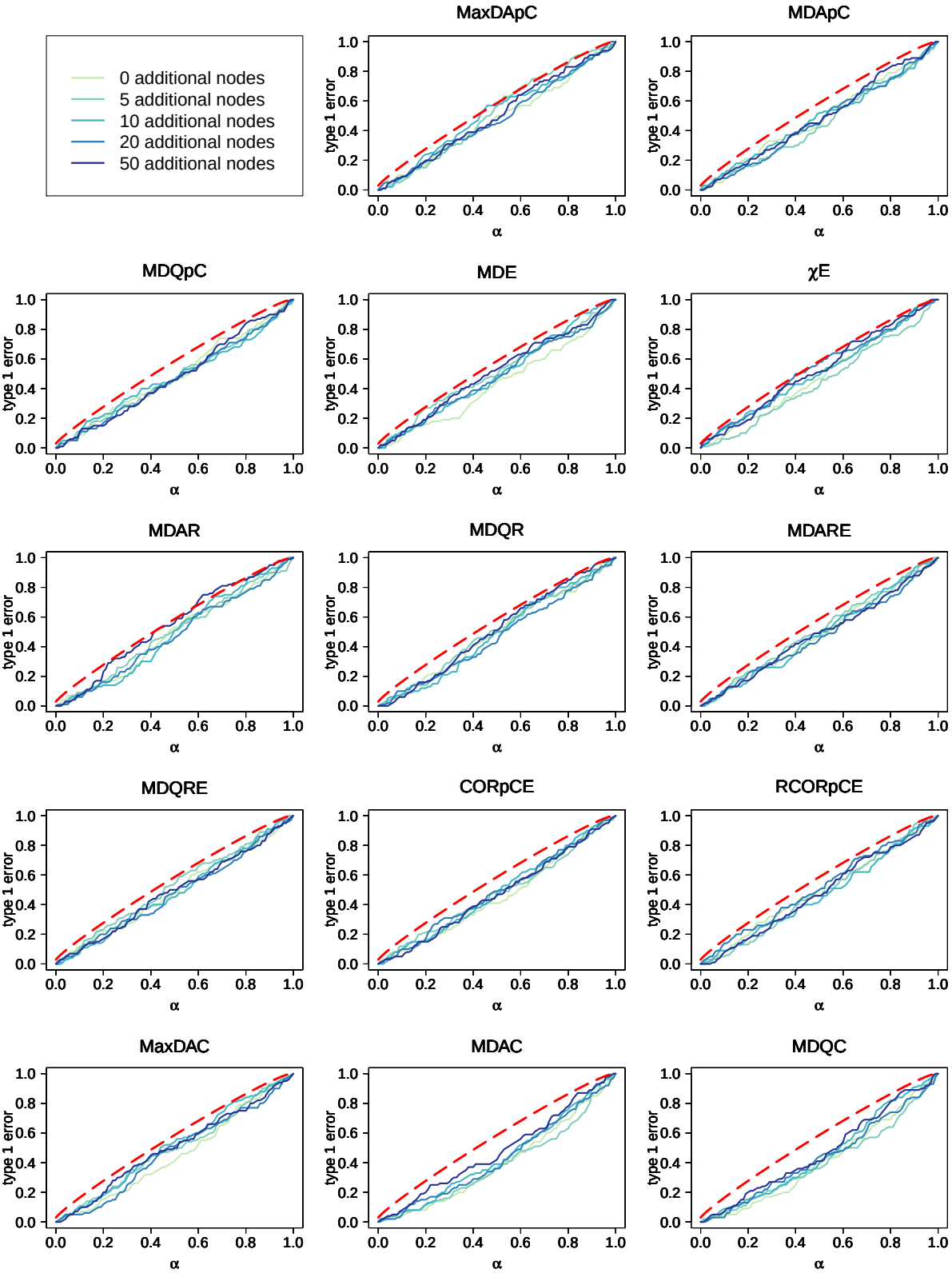

Figure 31: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 8.

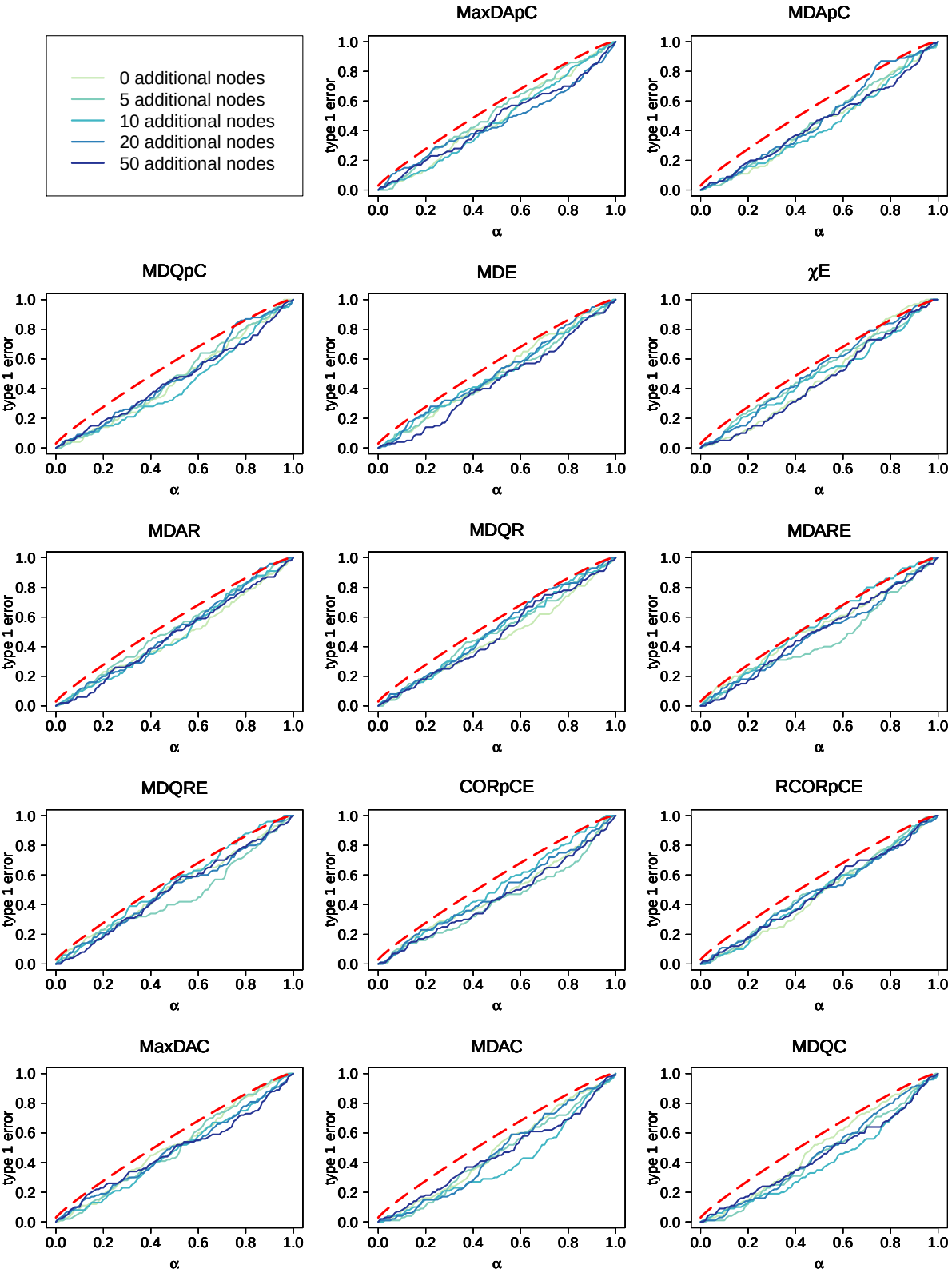

Figure 32: Proportion of misleadingly rejected hypothesis for simulated setting of 100 samples in each group and noise 16.

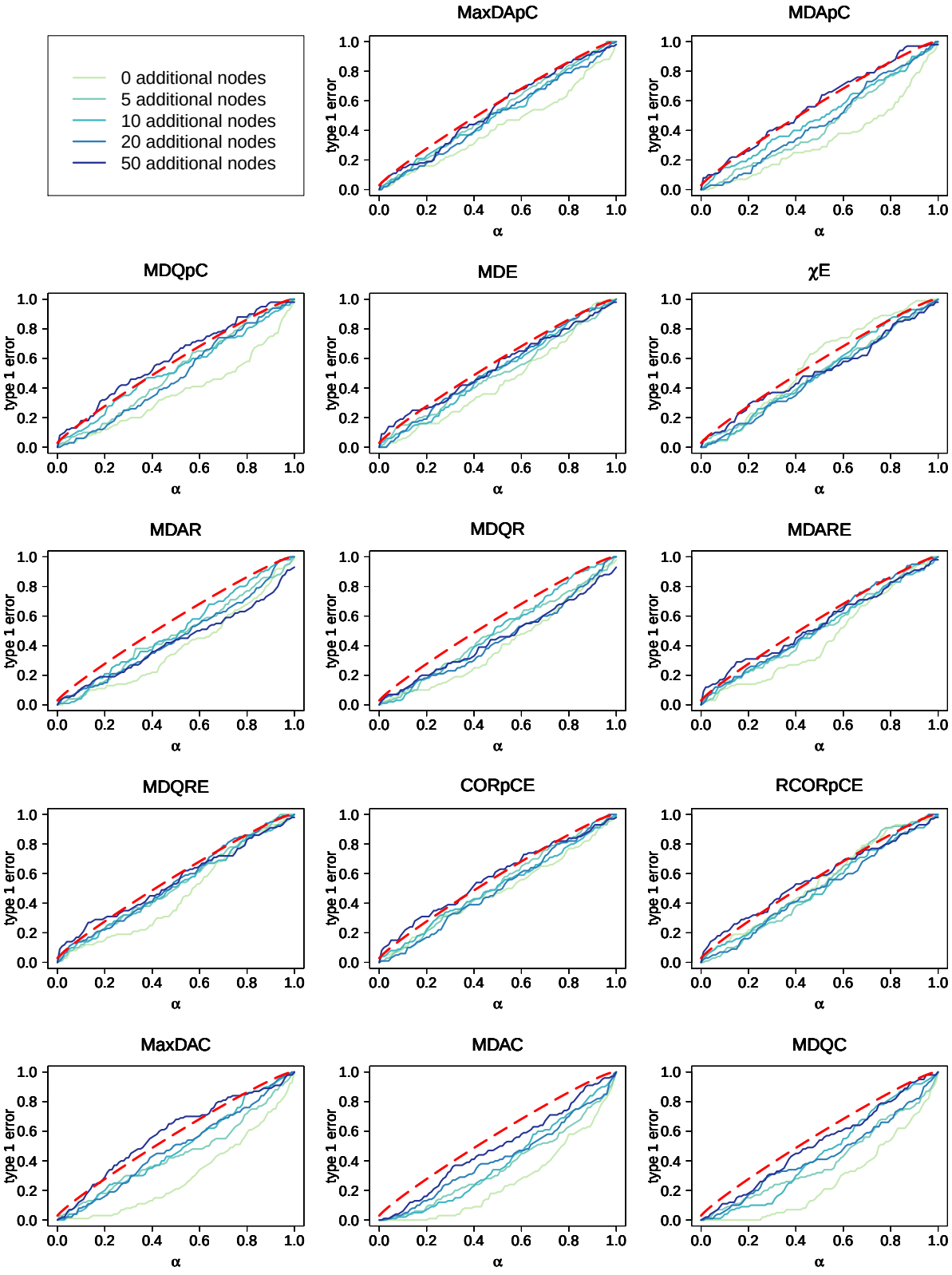

Figure 33: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 0.01.

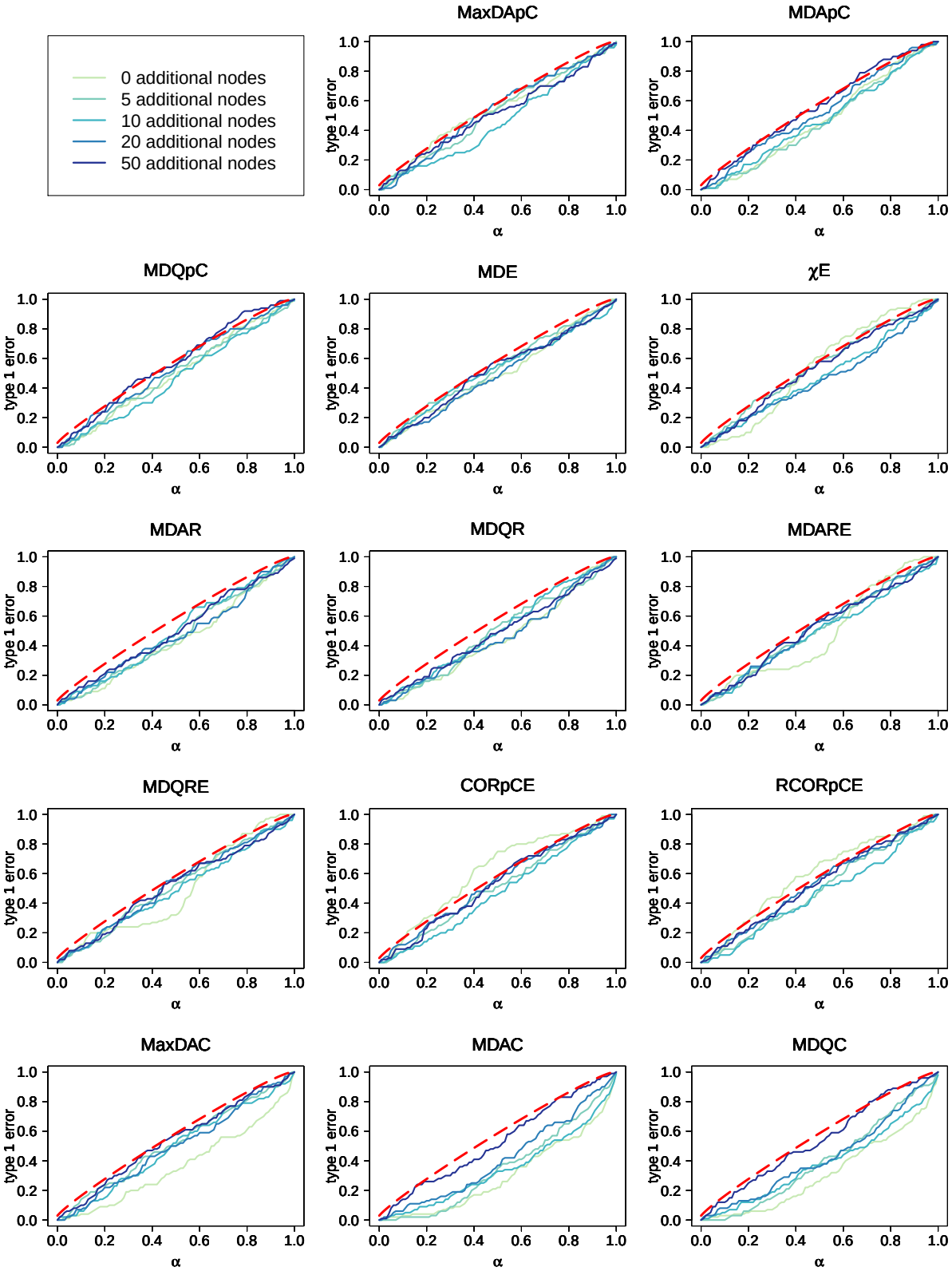

Figure 34: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 0.1.

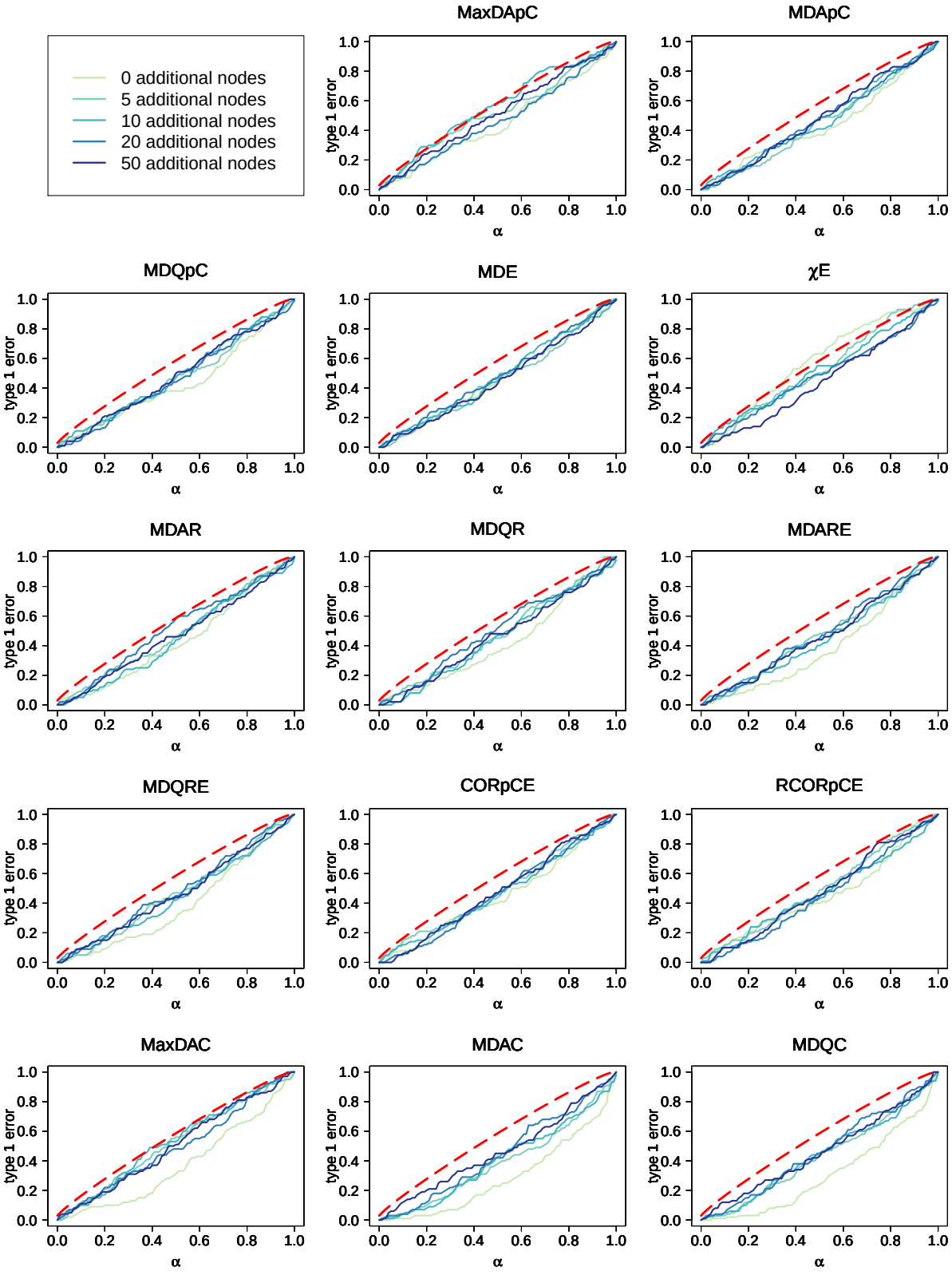

Figure 35: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 0.5.

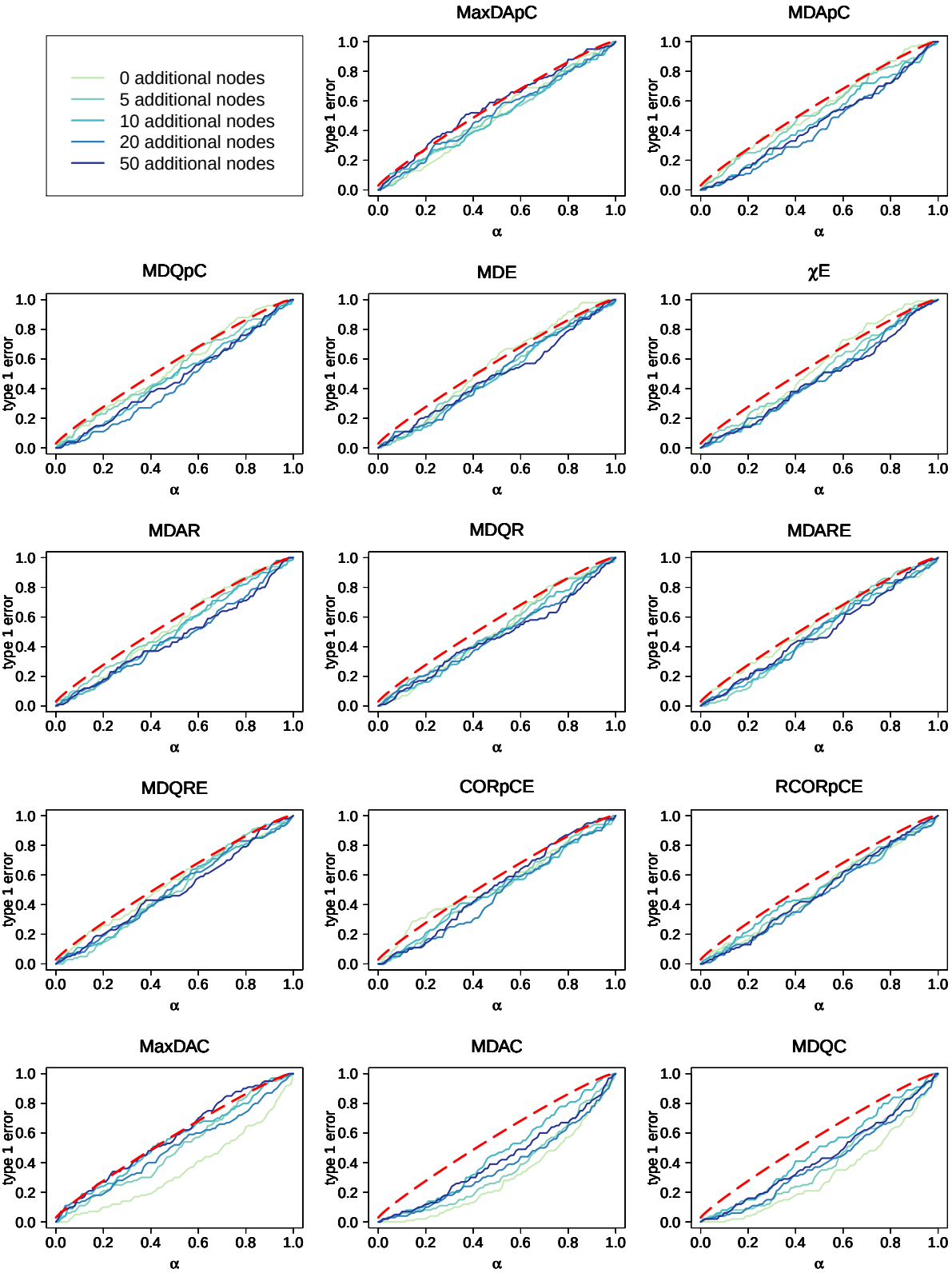

Figure 36: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 1.

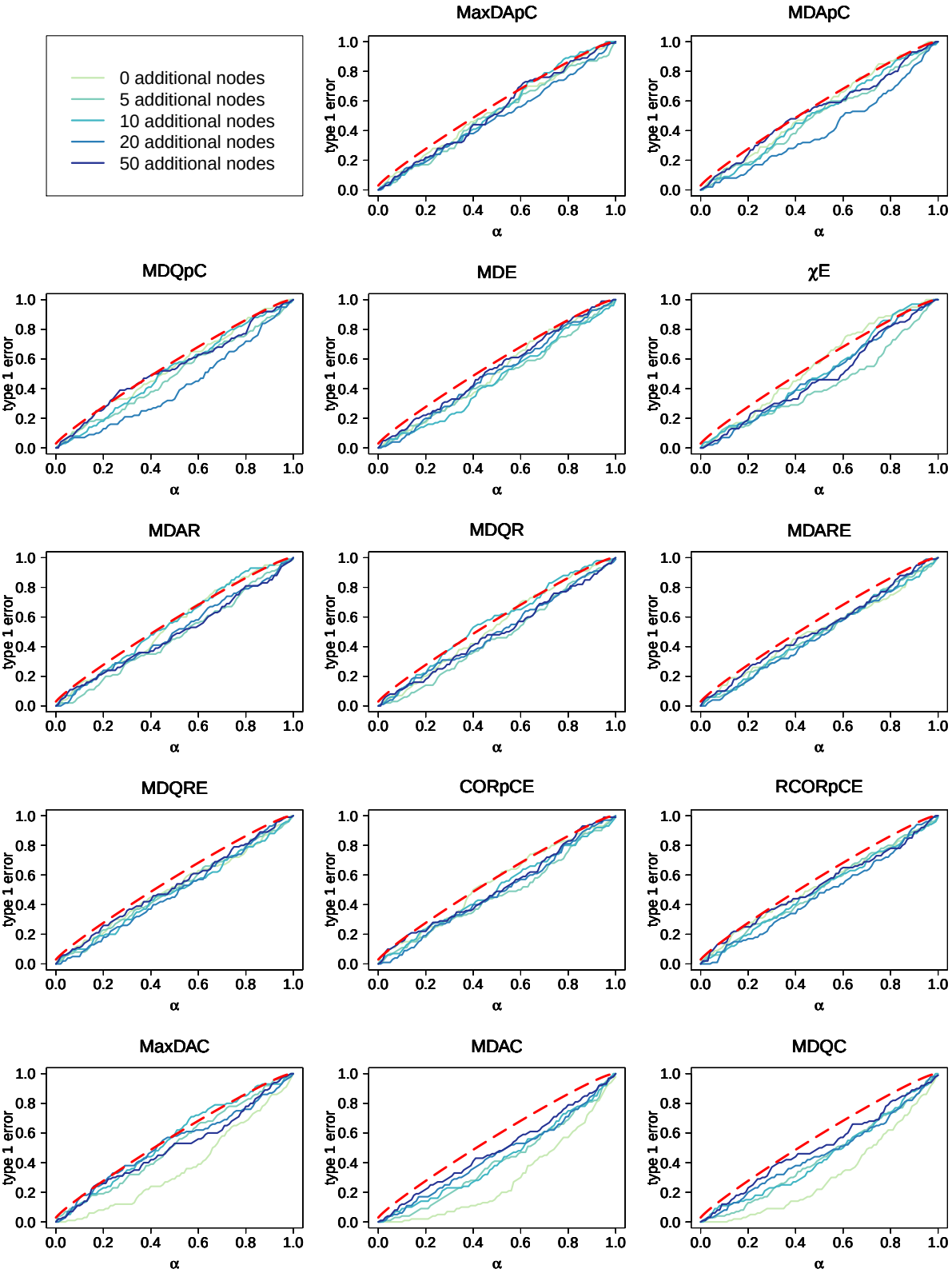

Figure 37: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 2.

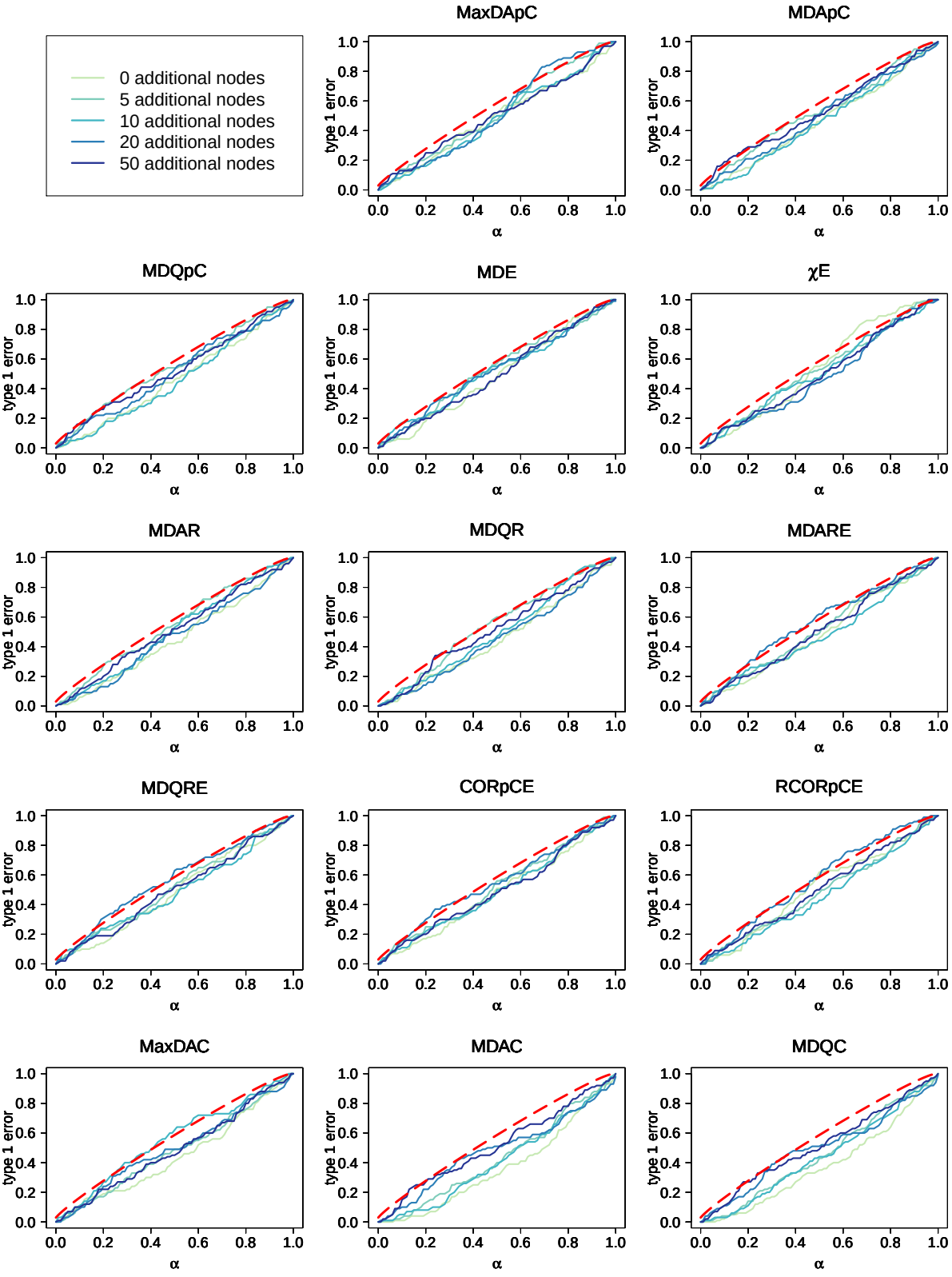

Figure 38: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 4.

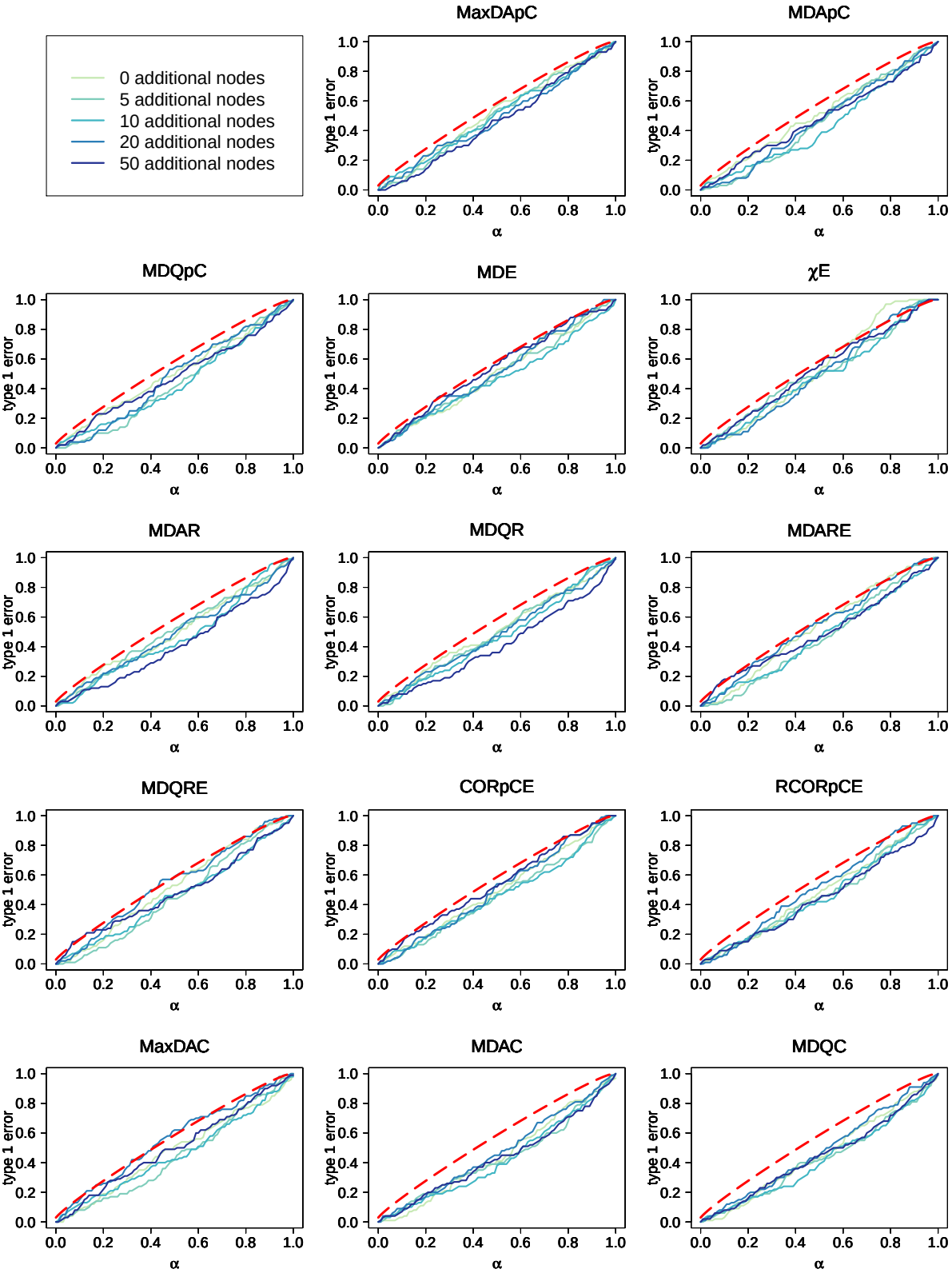

Figure 39: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 8.

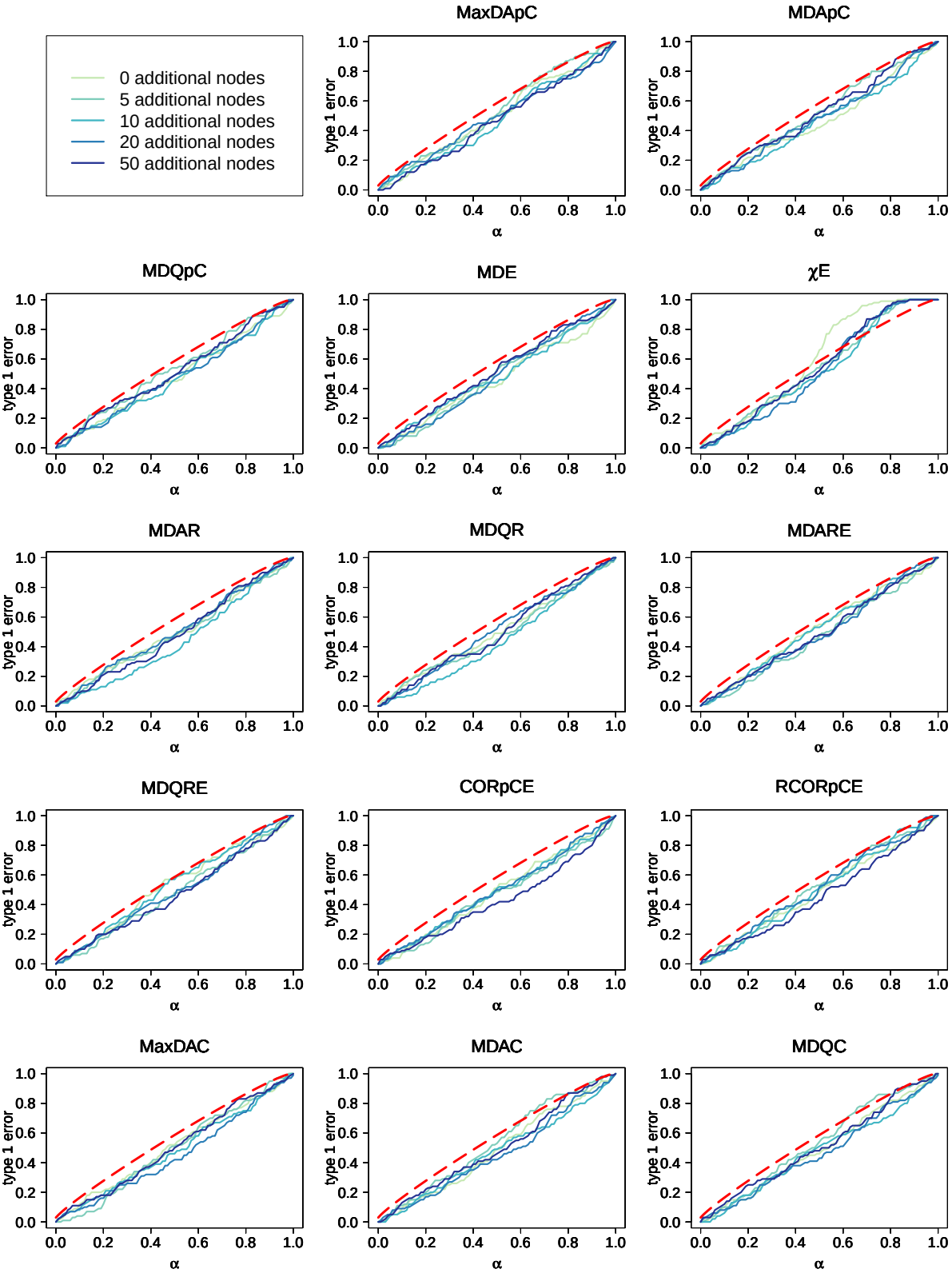

Figure 40: Proportion of misleadingly rejected hypothesis for simulated setting of 150 and 50 samples per group and noise 16.

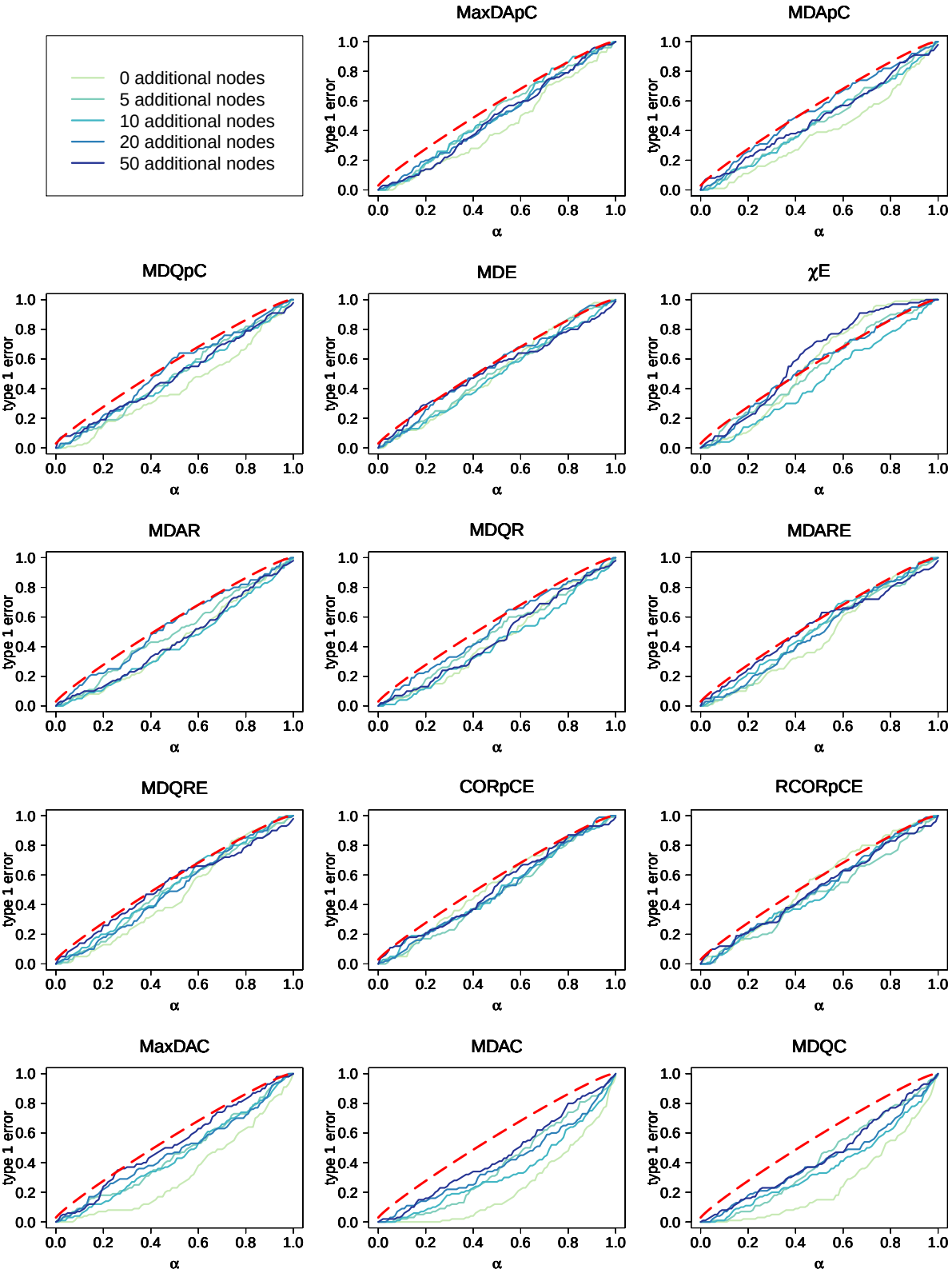

Figure 41: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 0.01.

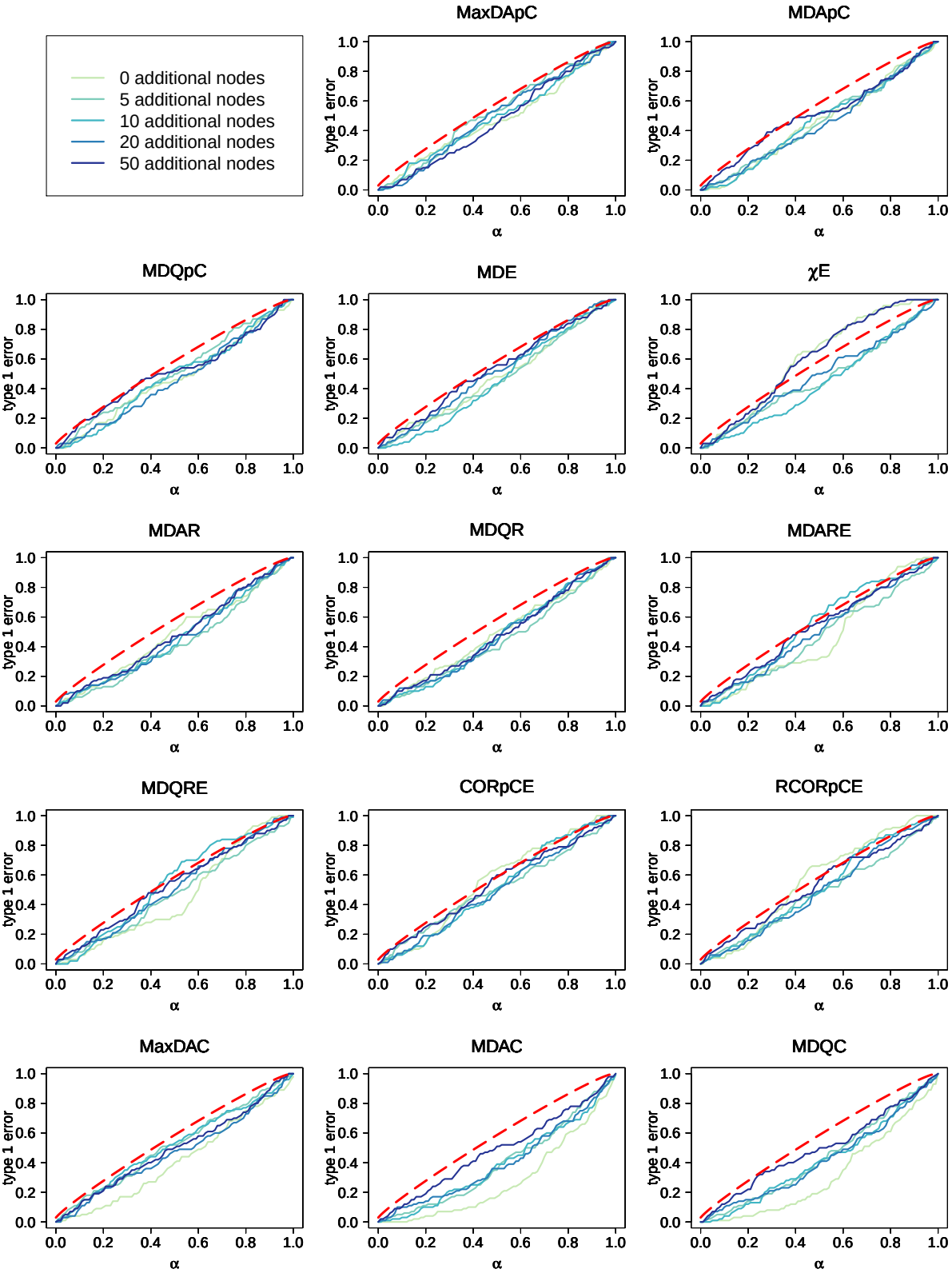

Figure 42: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 0.1.

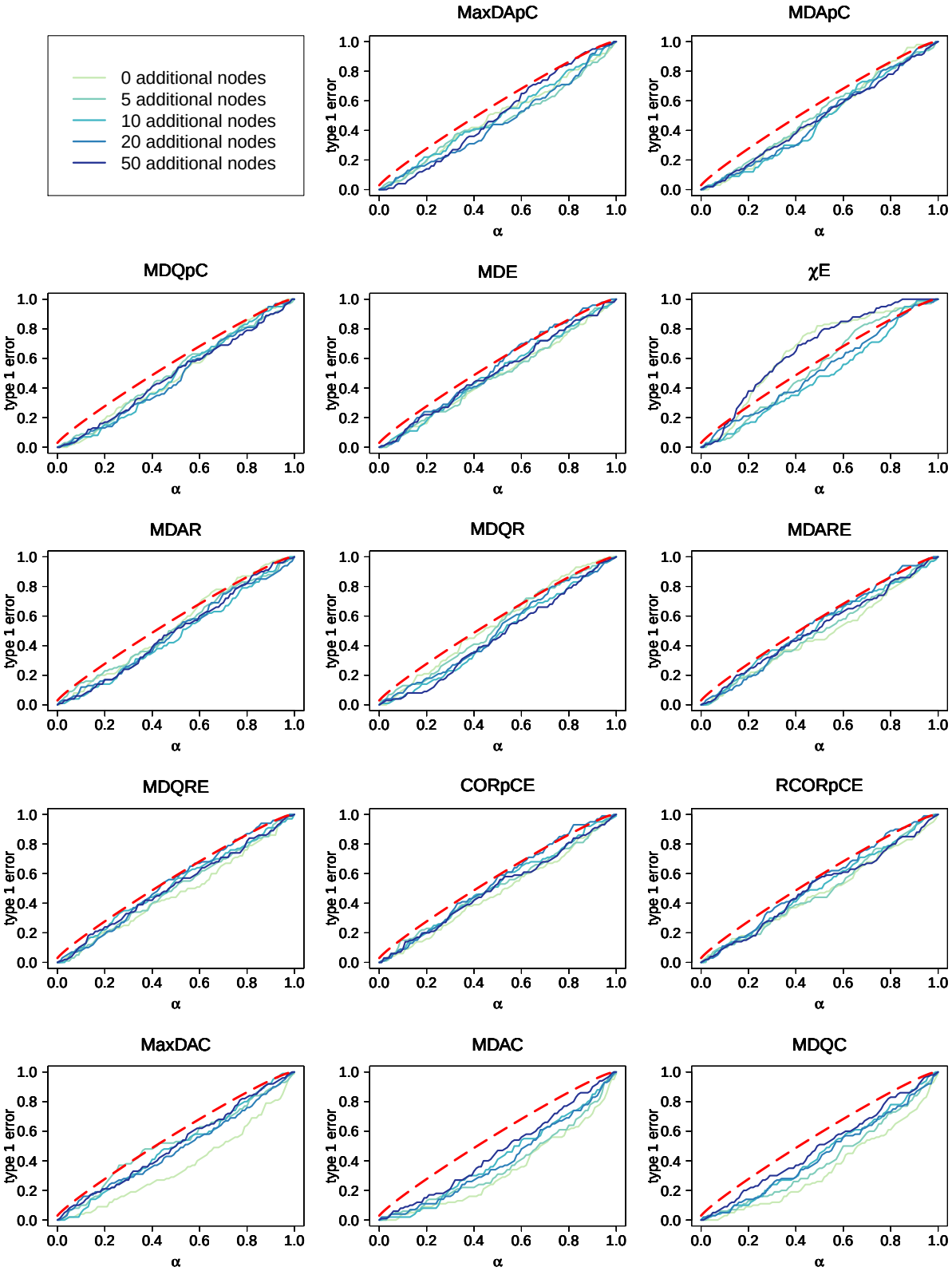

Figure 43: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 0.5.

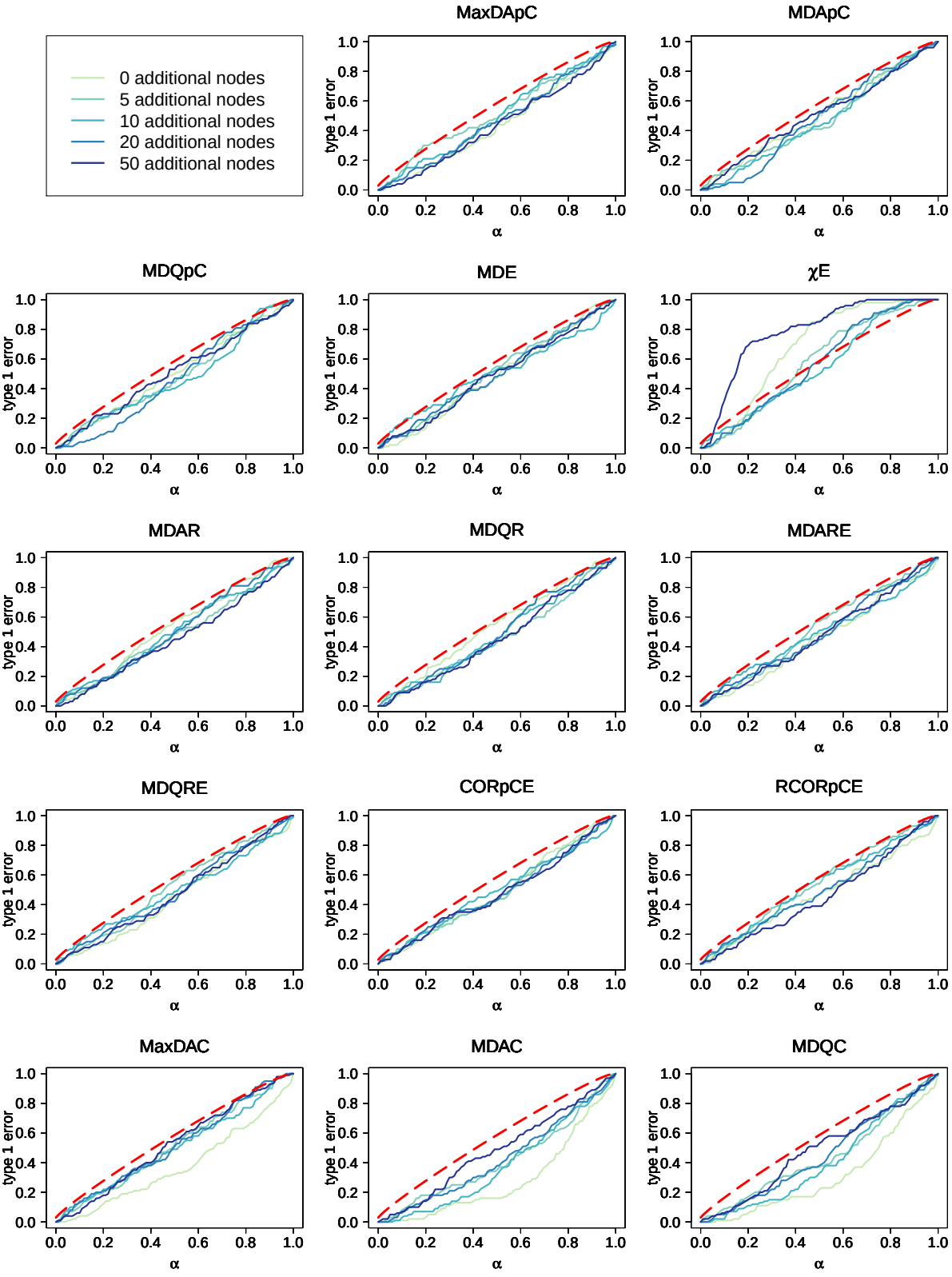

Figure 44: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 1.

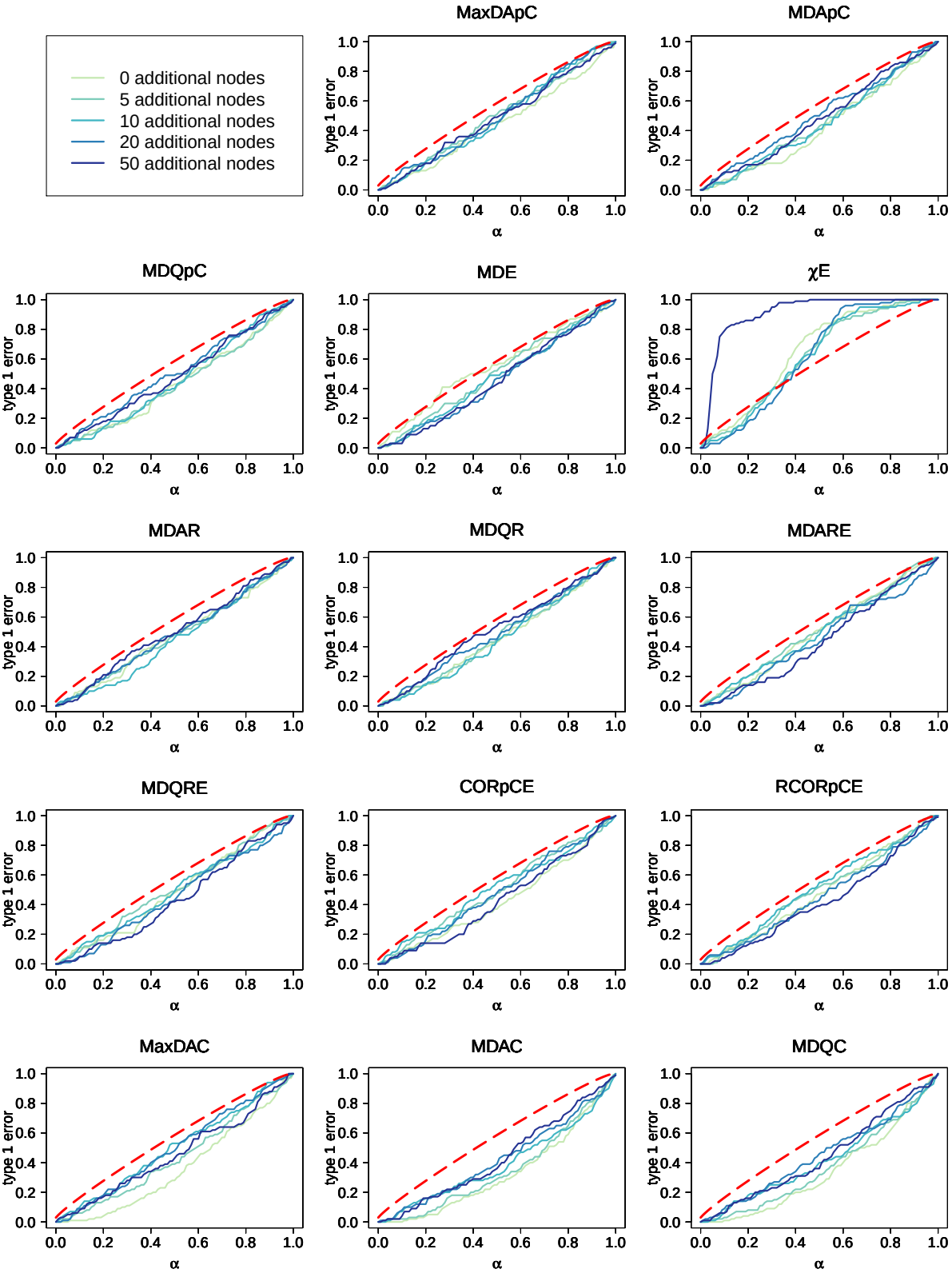

Figure 45: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 2.

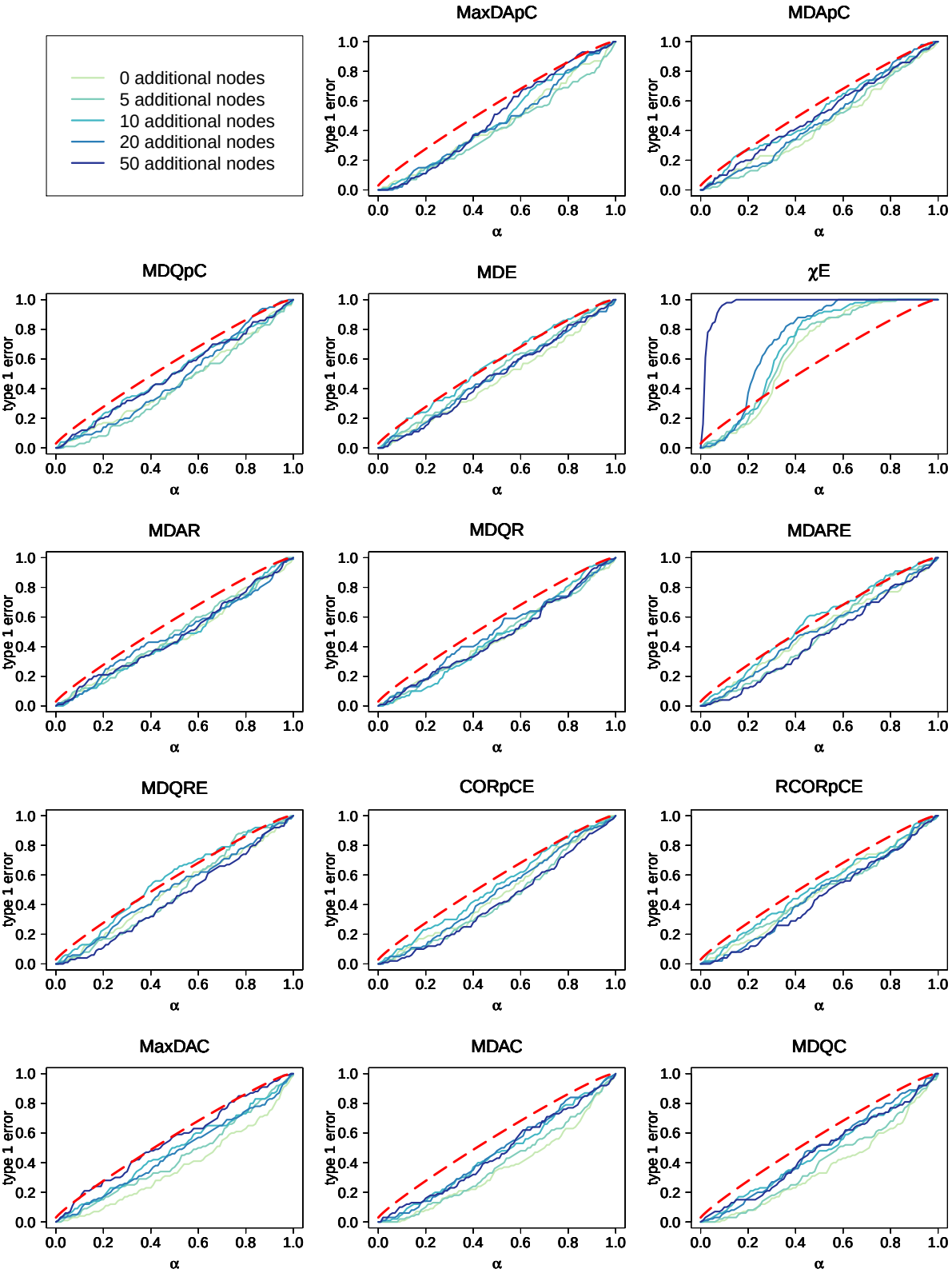

Figure 46: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 4.

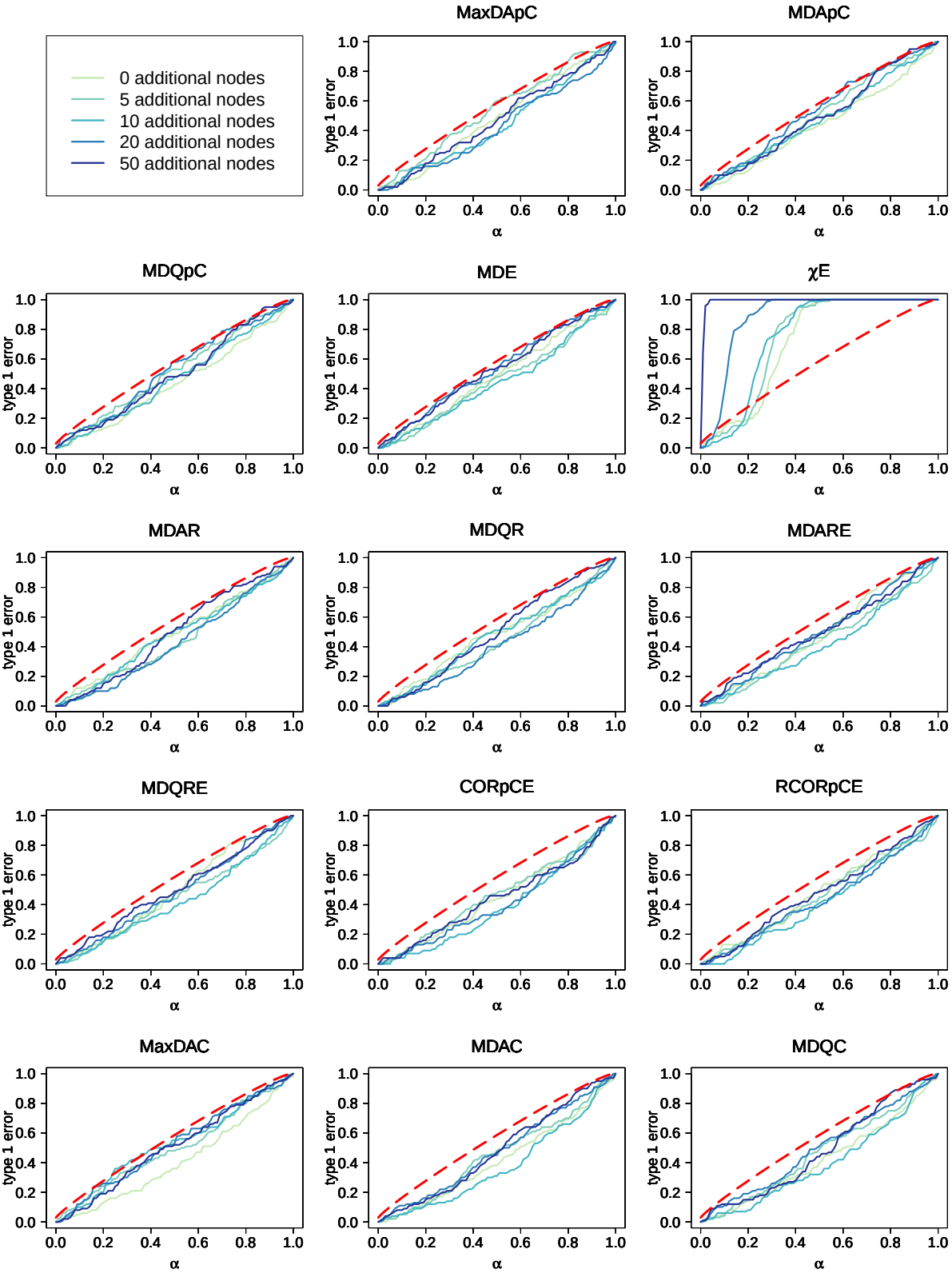

Figure 47: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 4.

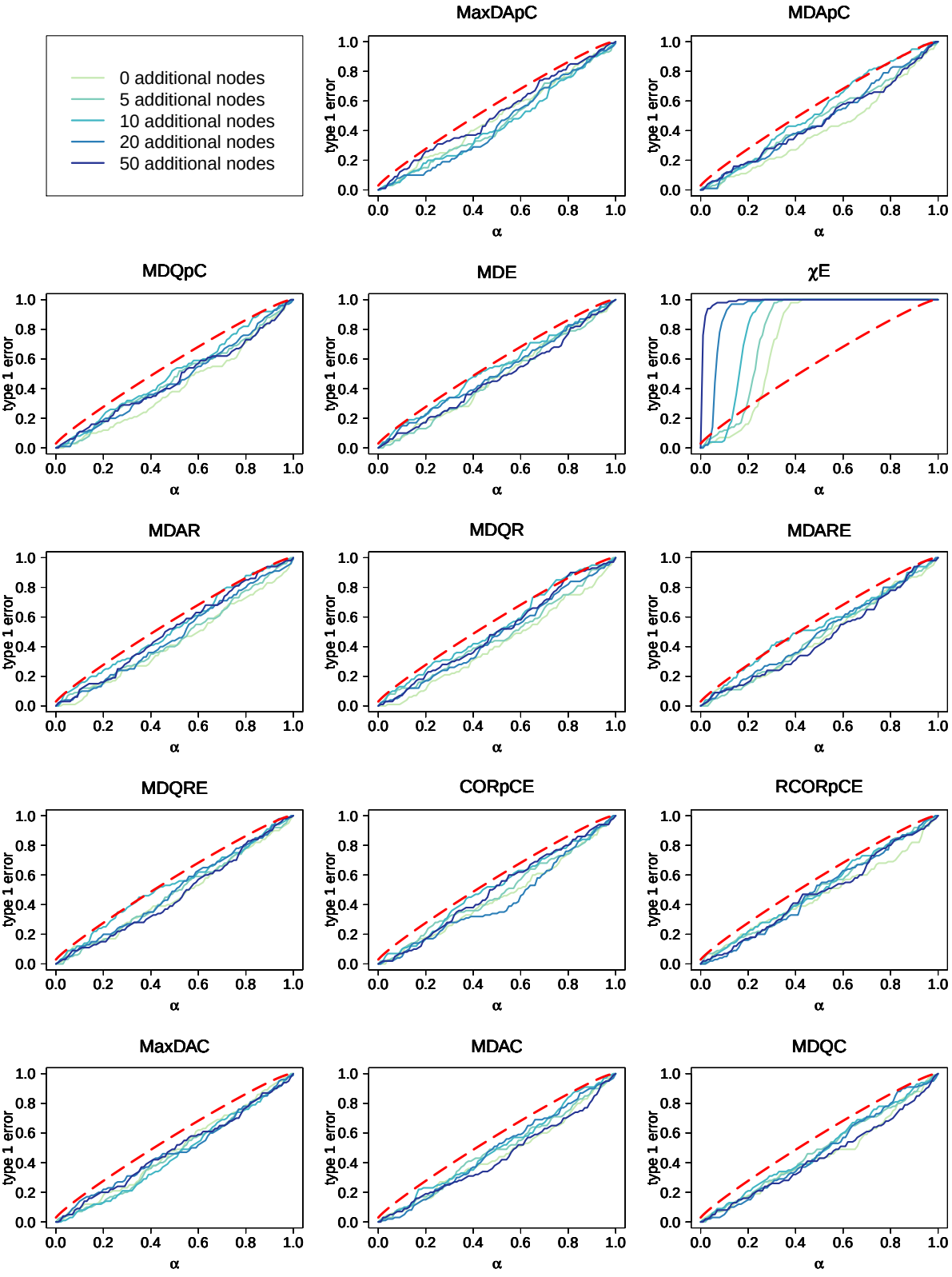

Figure 48: Proportion of misleadingly rejected hypothesis for simulated setting of 180 and 20 samples per group and noise 16.

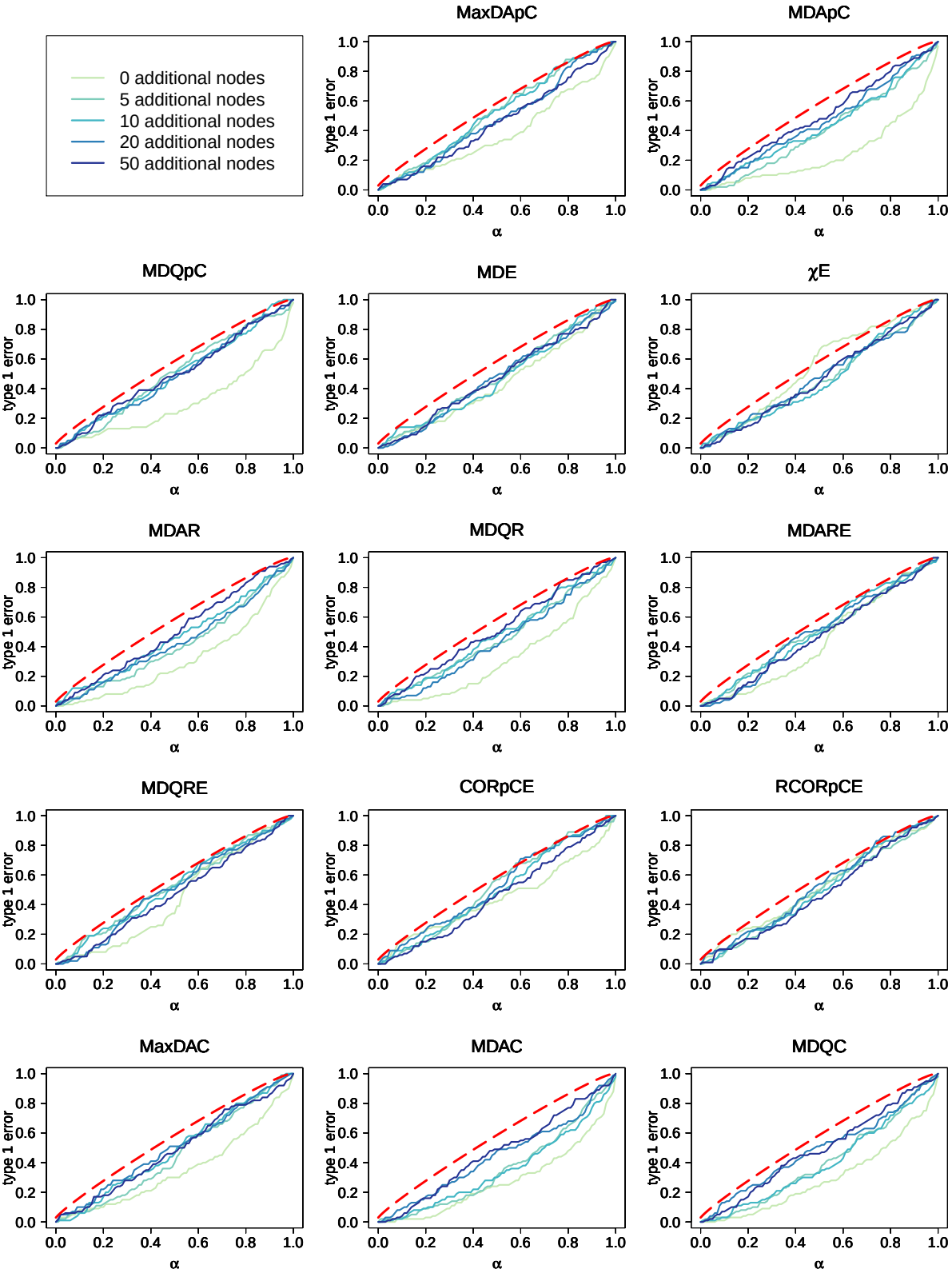

Figure 49: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 0.01.

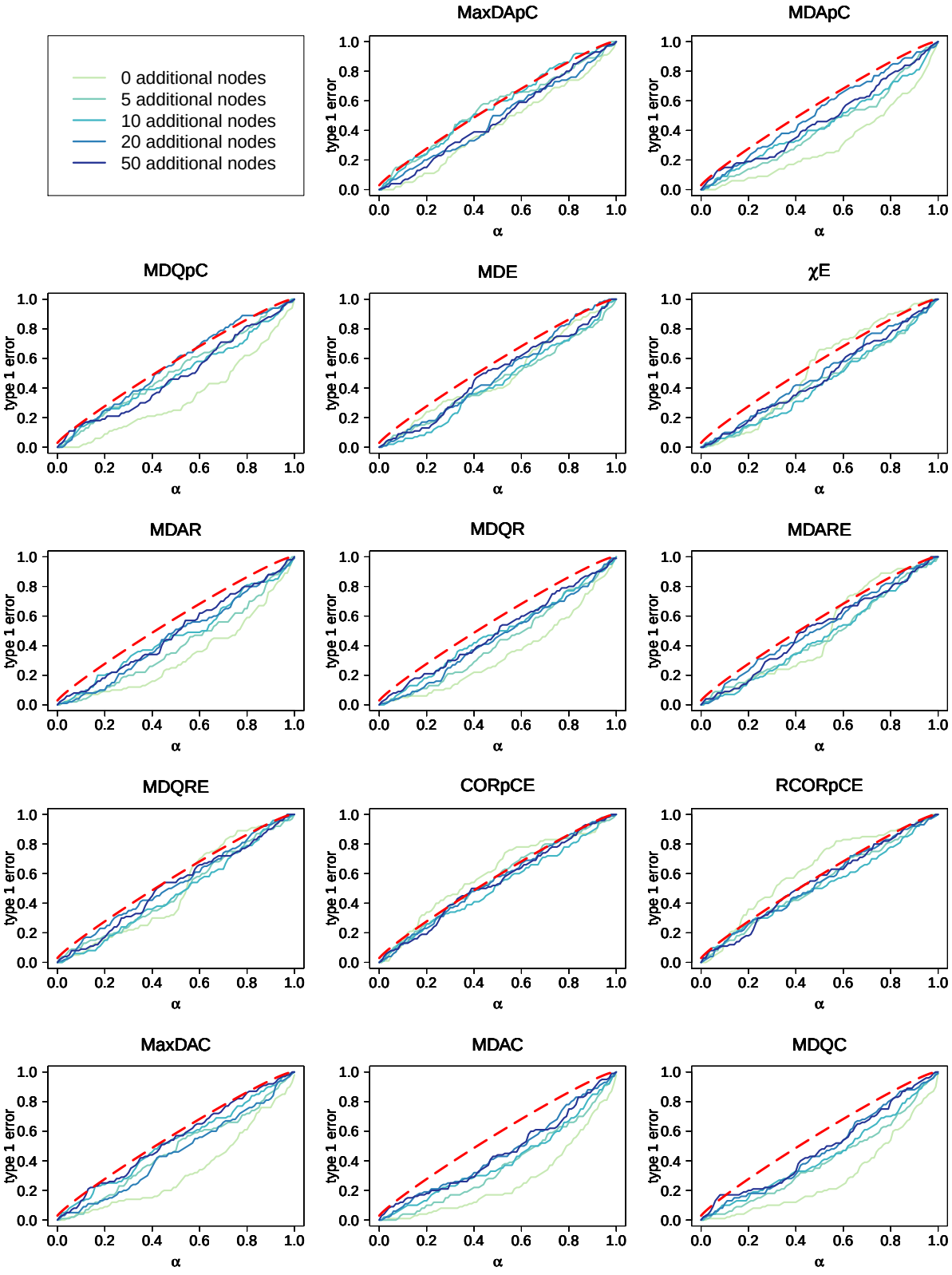

Figure 50: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 0.1.

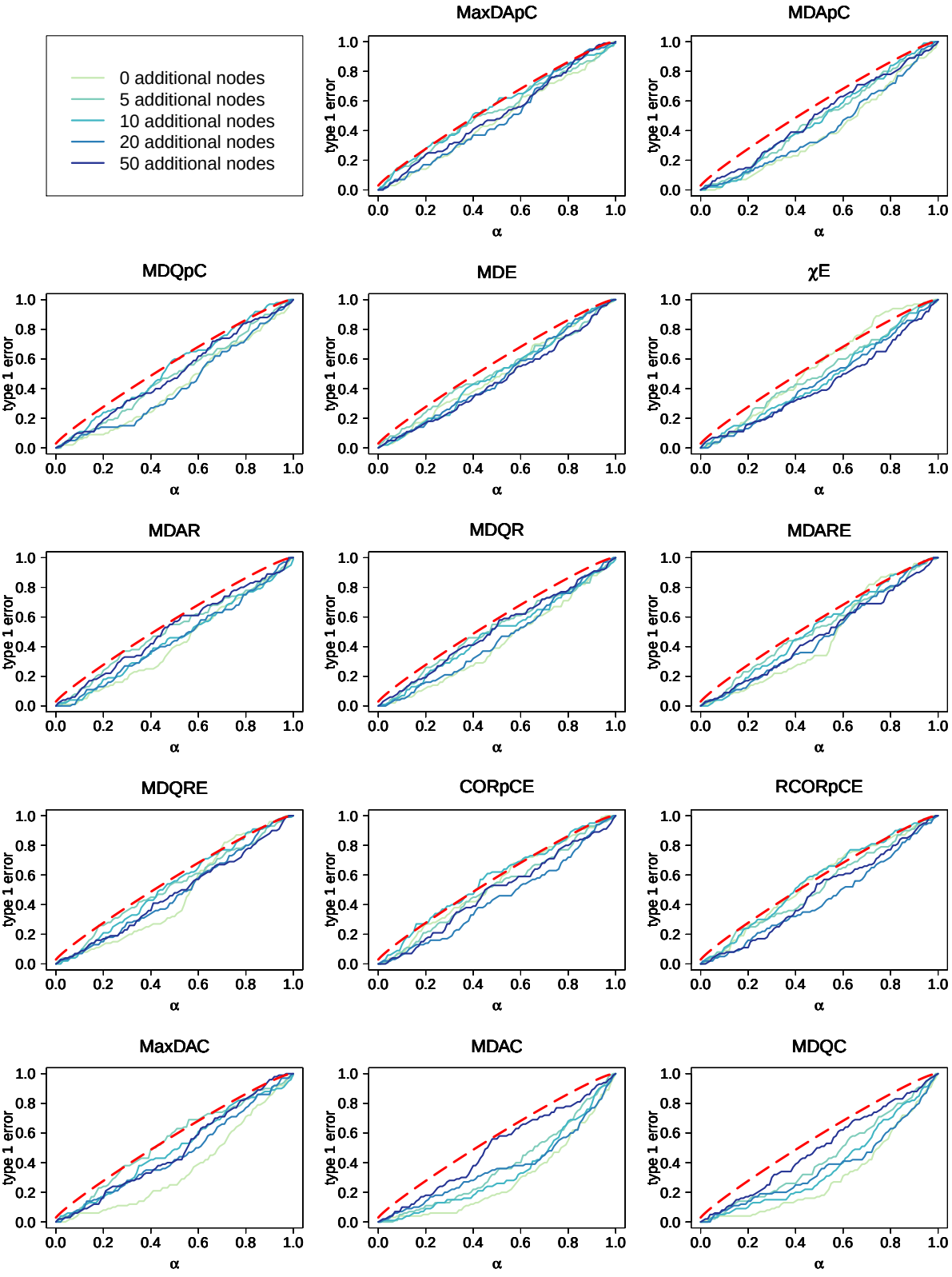

Figure 51: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 0.5.

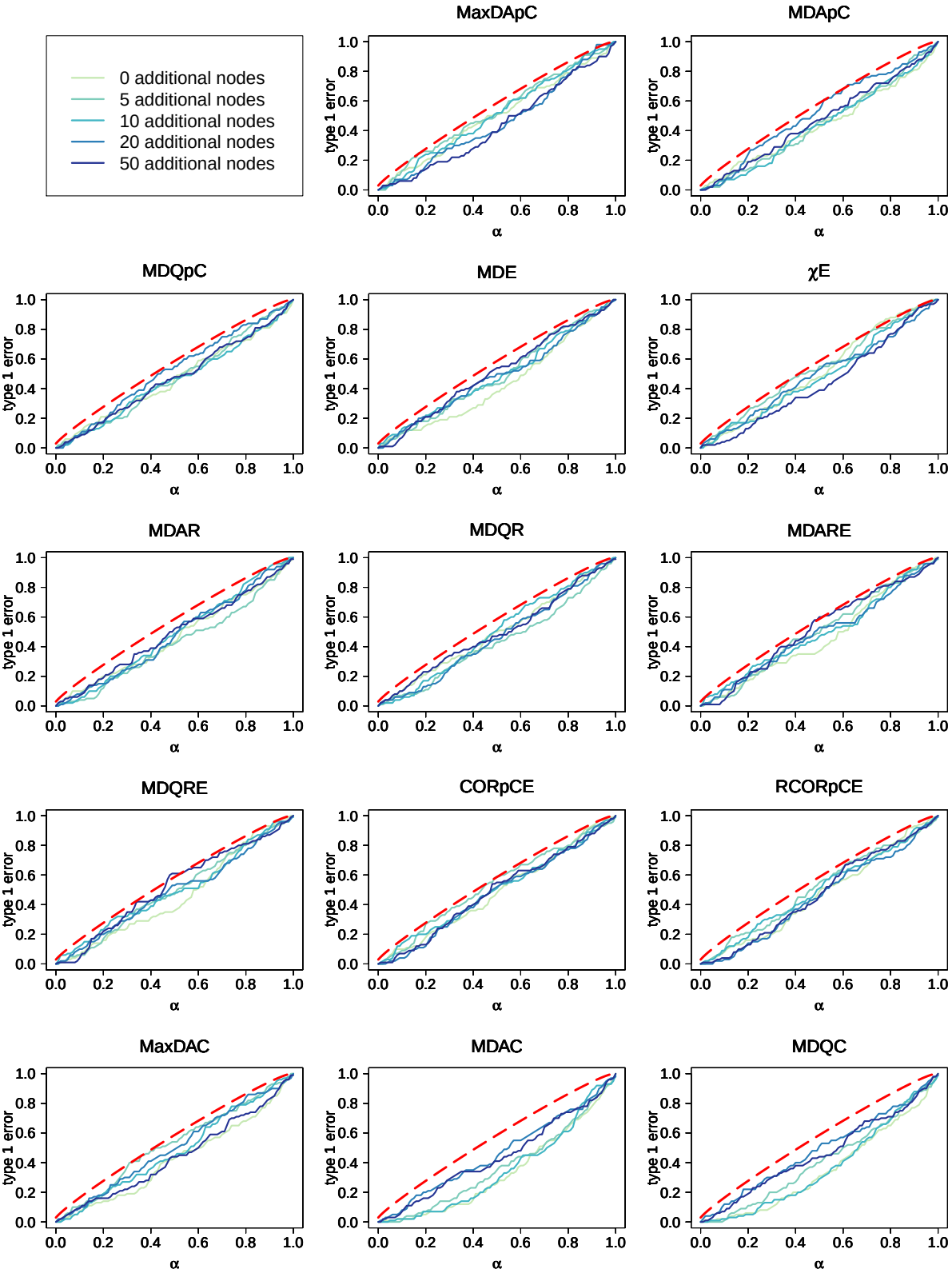

Figure 52: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 1.

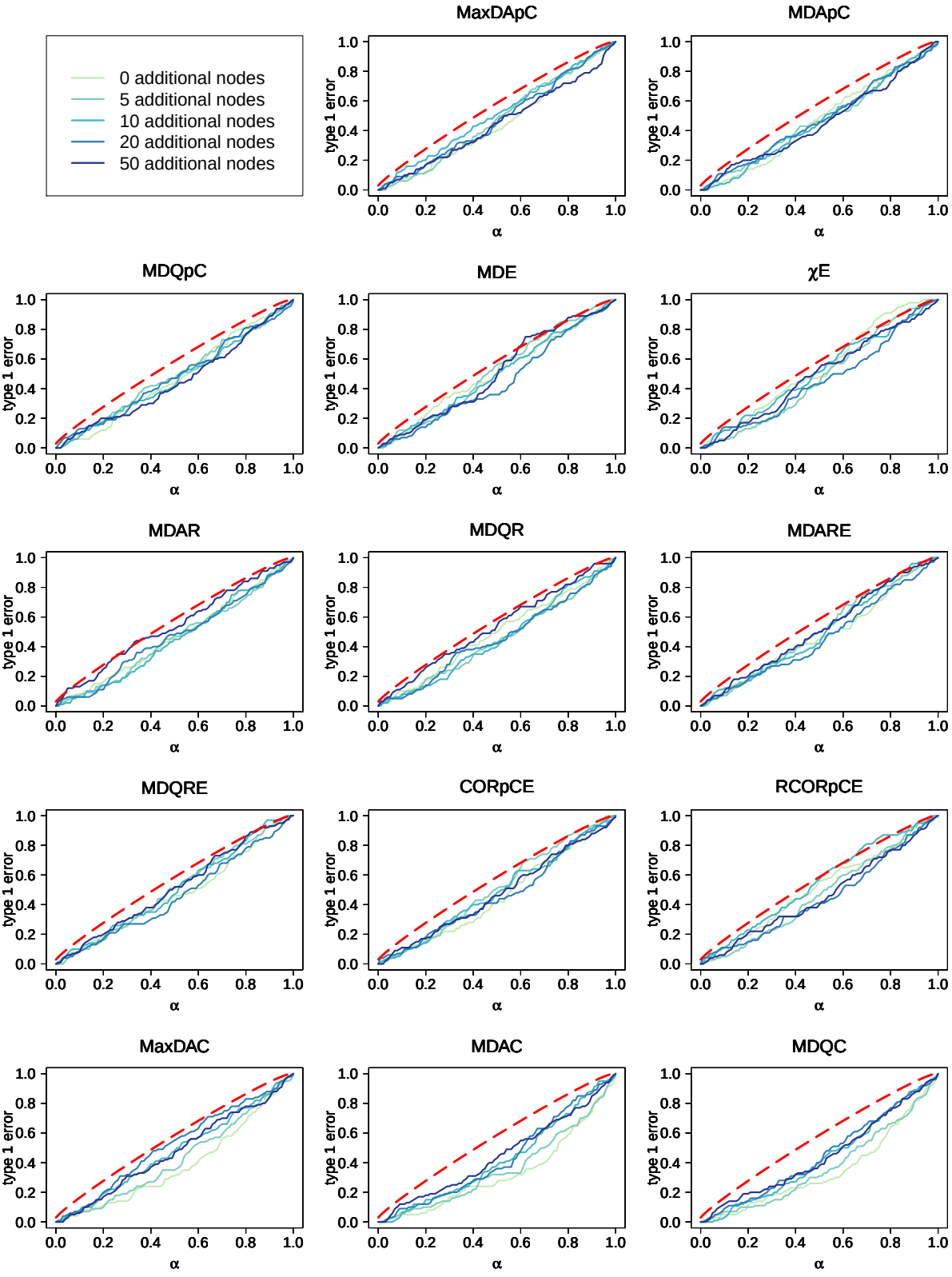

Figure 53: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 2.

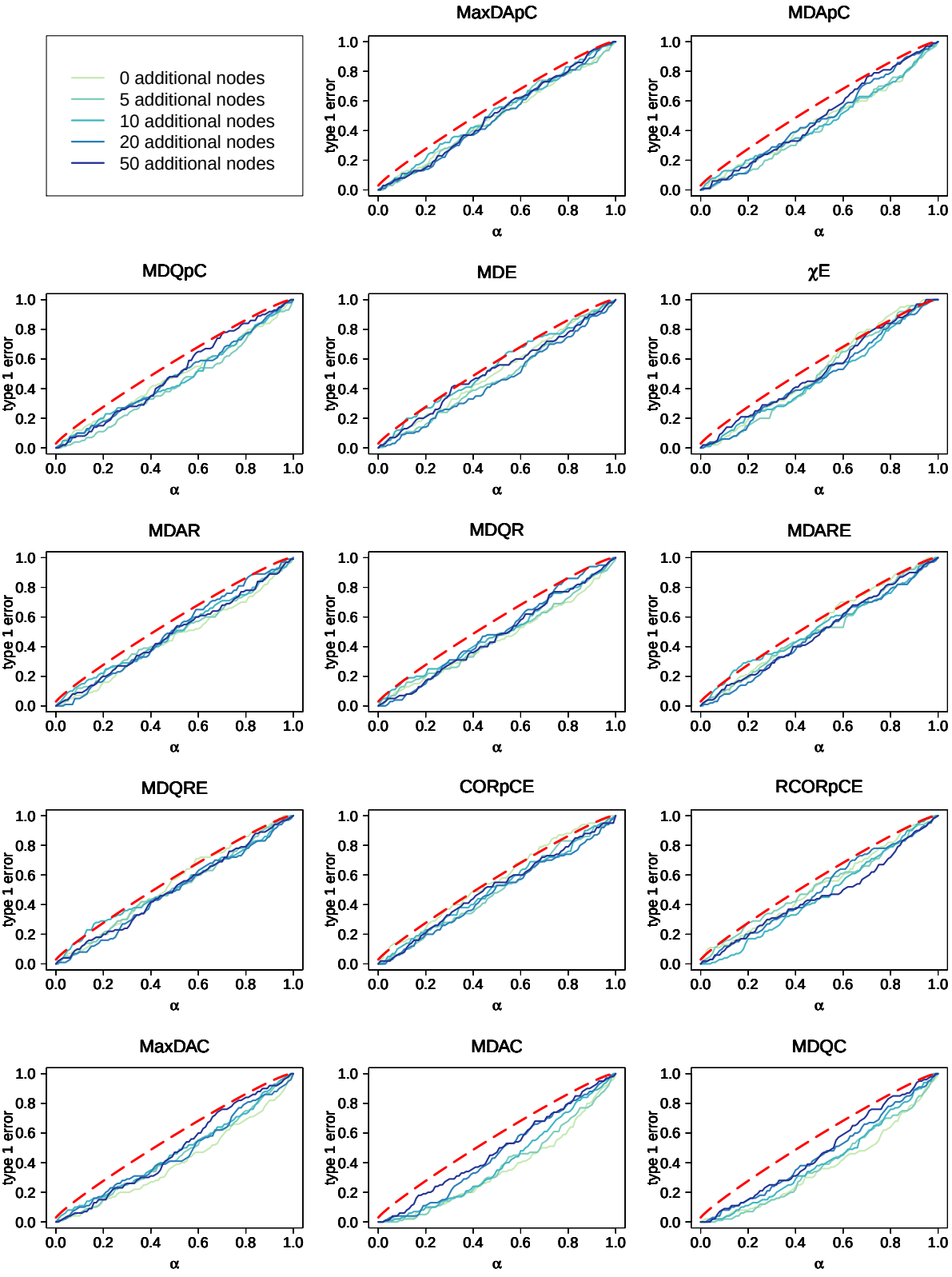

Figure 54: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 4.

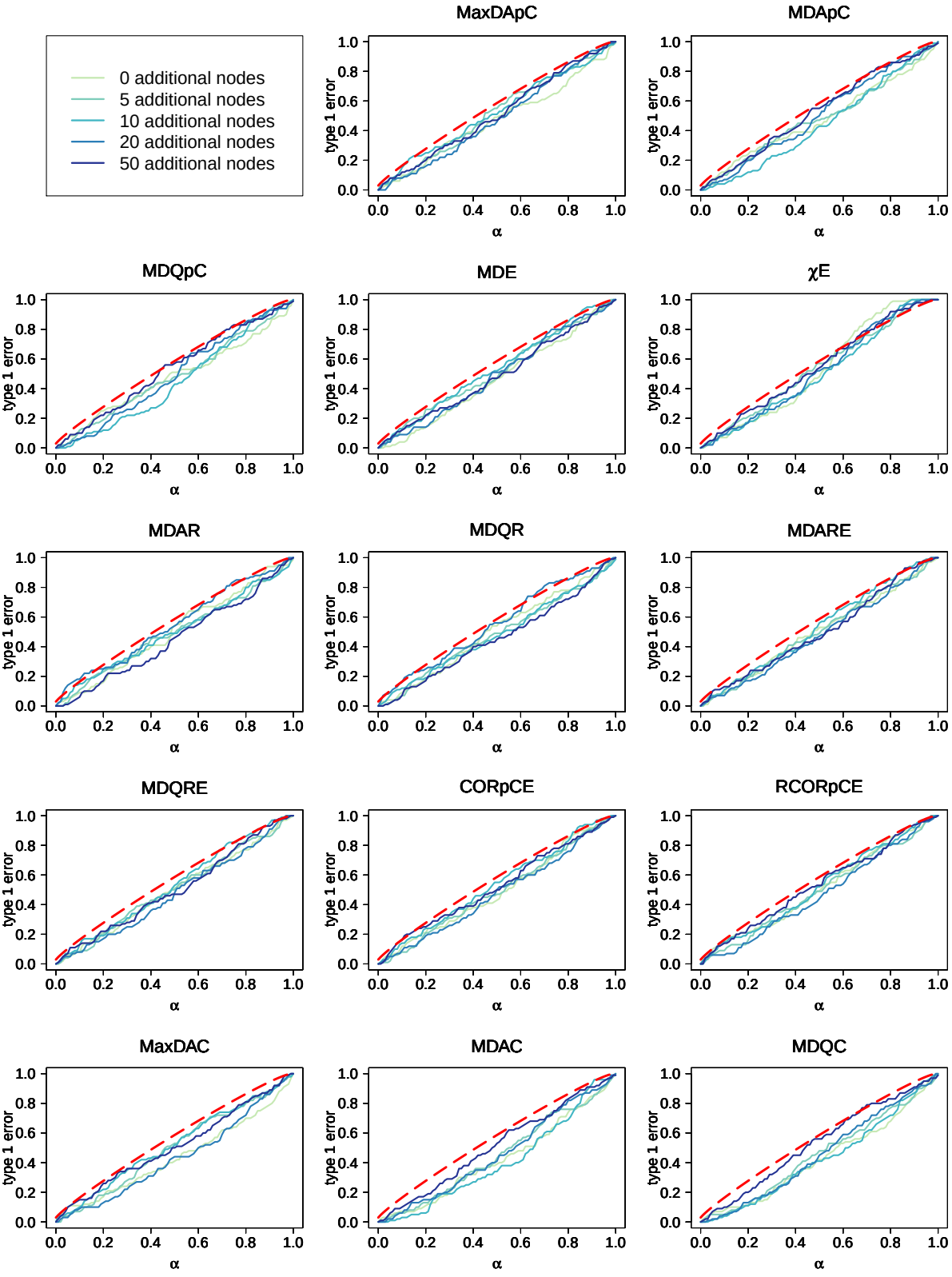

Figure 55: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 8.

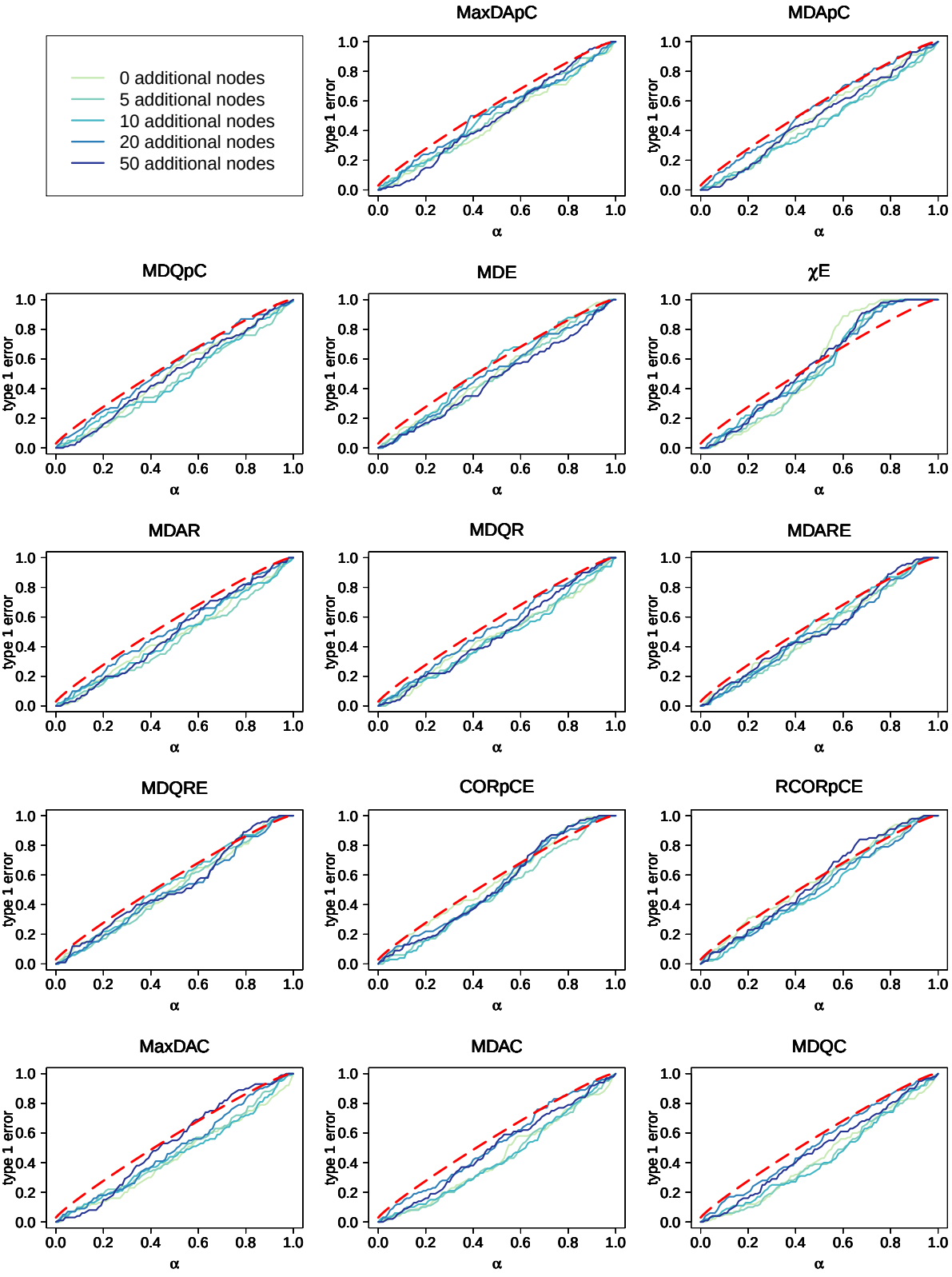

Figure 56: Proportion of misleadingly rejected hypothesis for simulated setting of 50 samples in each group and noise 16.

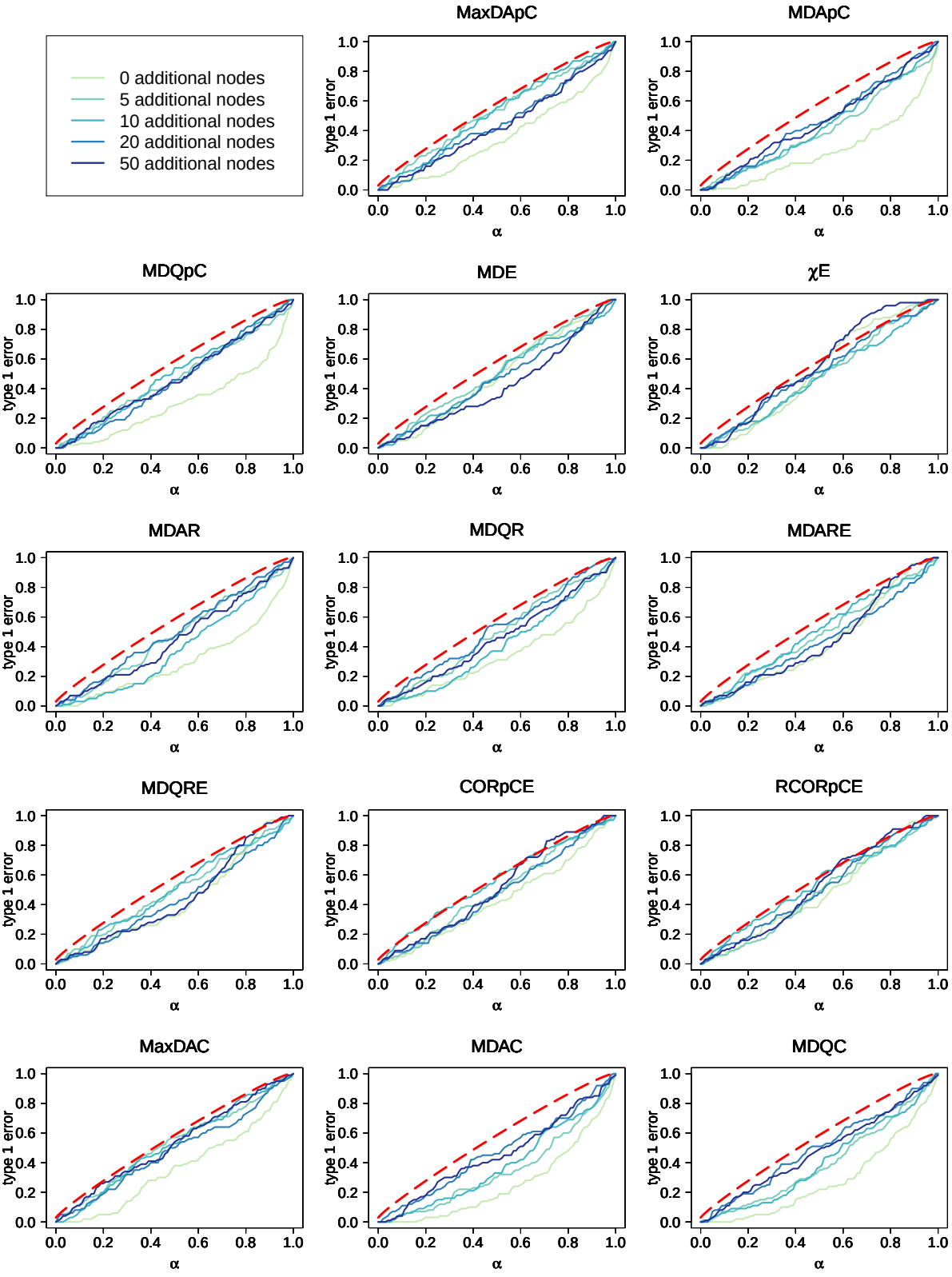

Figure 57: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 0.01.

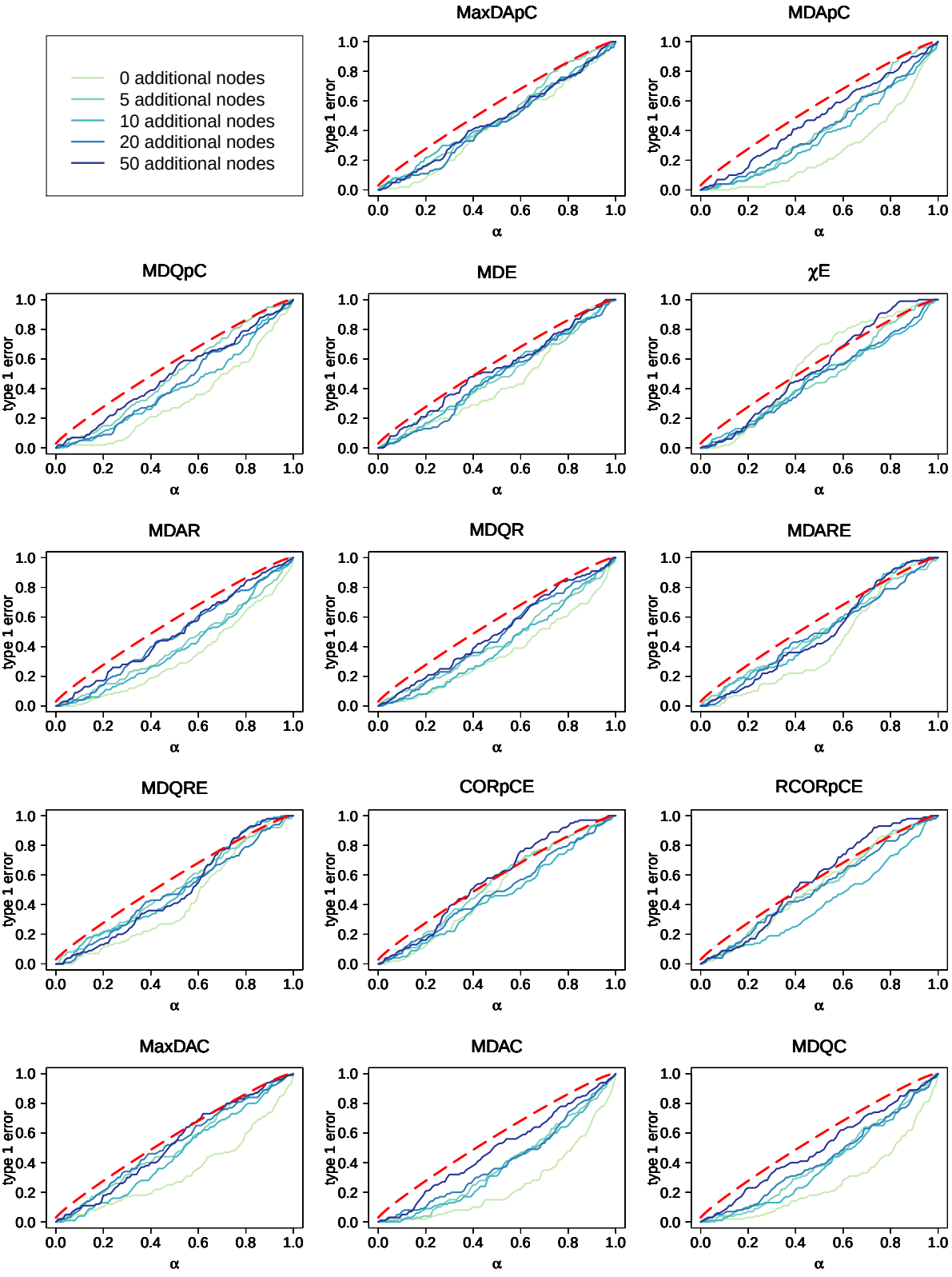

Figure 58: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 0.1.

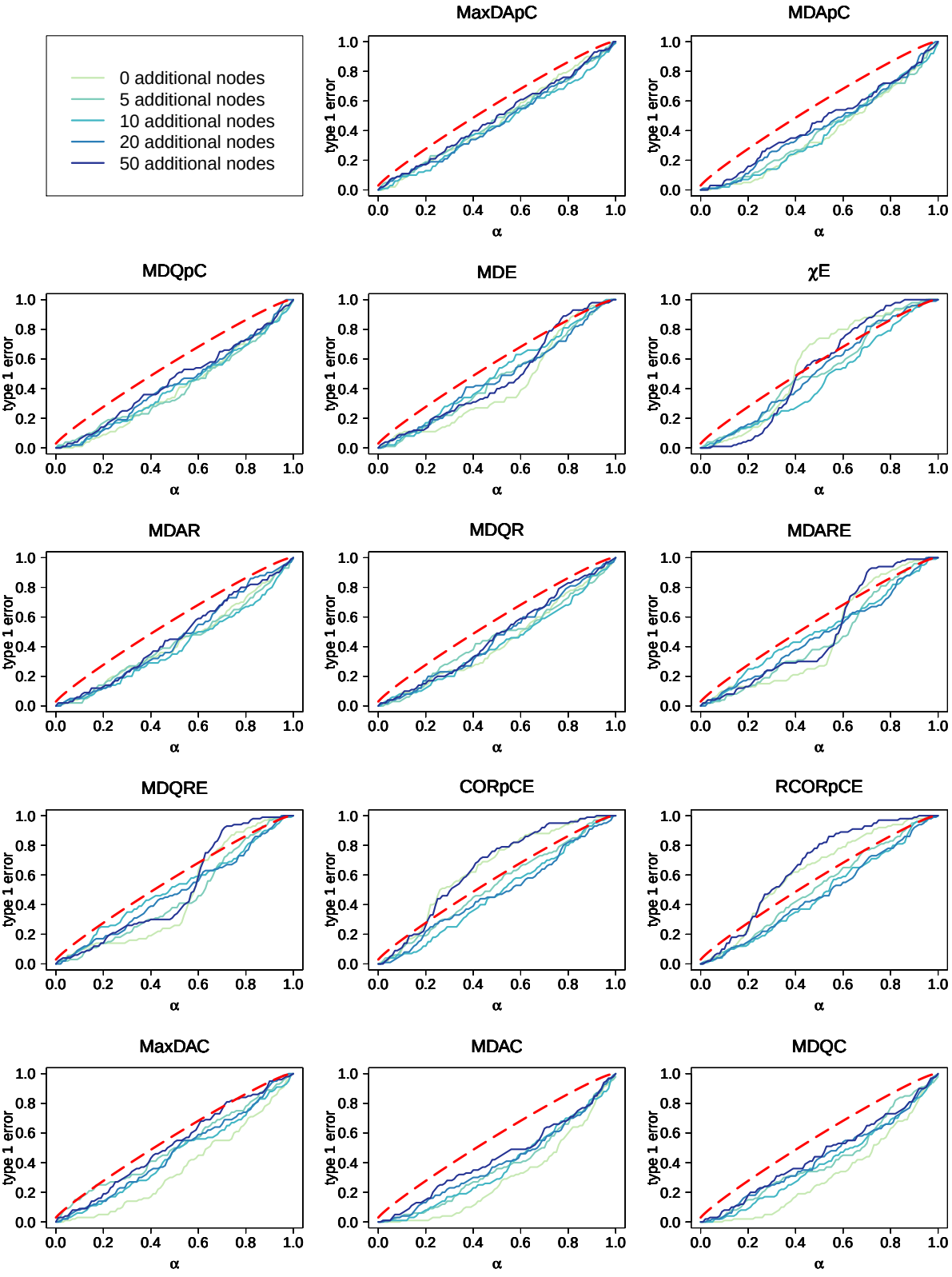

Figure 59: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 0.5.

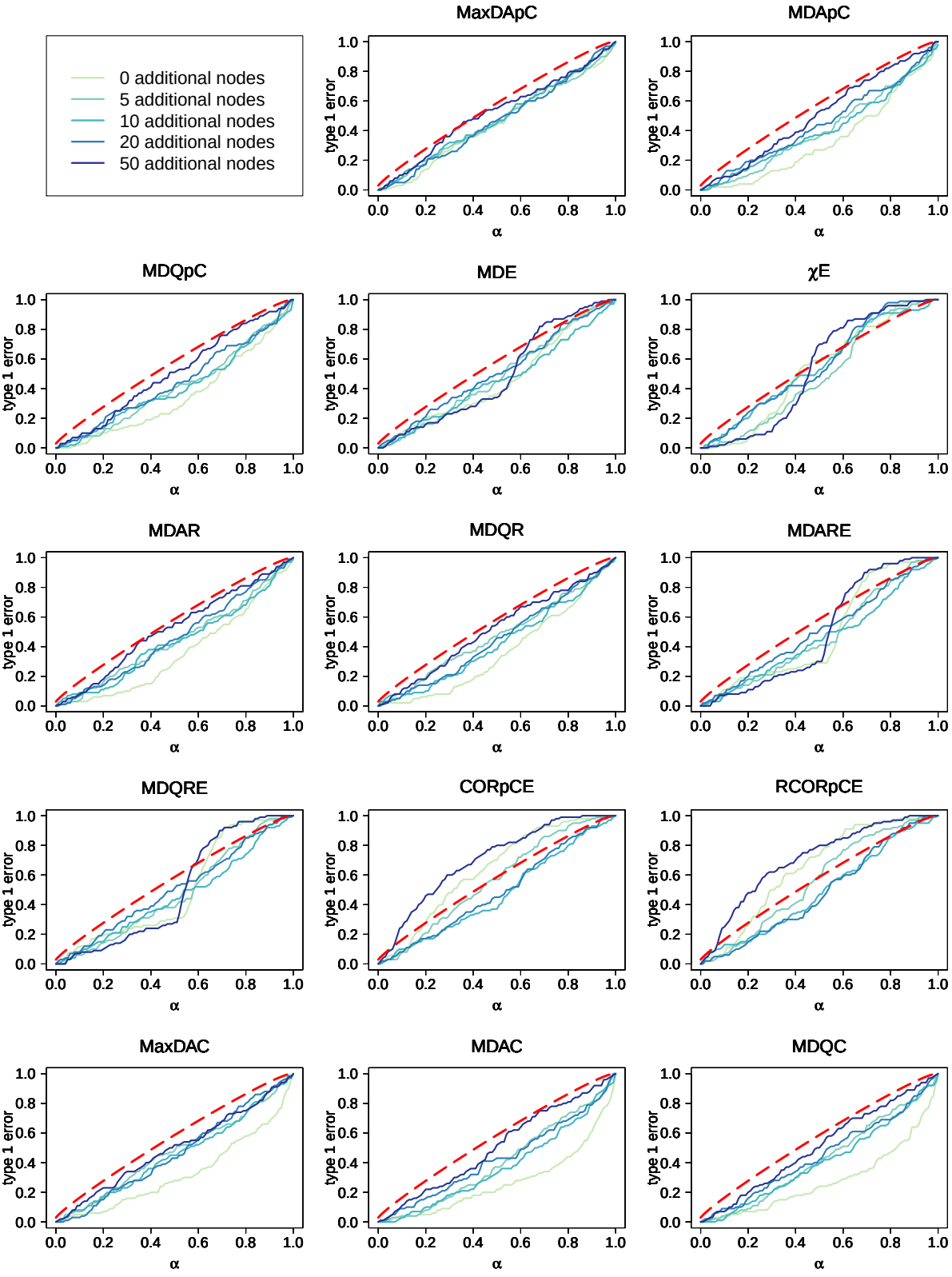

Figure 60: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 1.

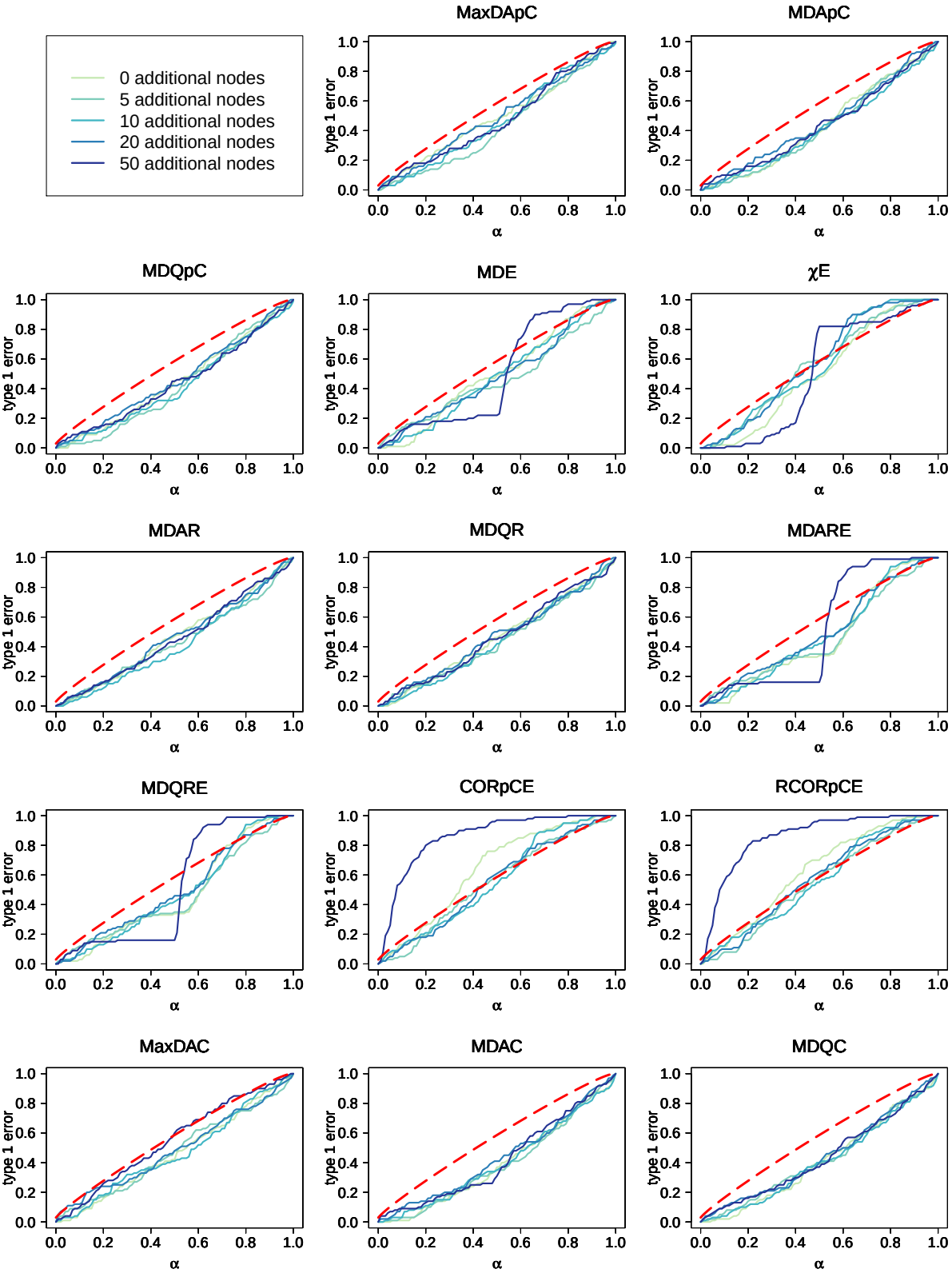

Figure 61: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 2.

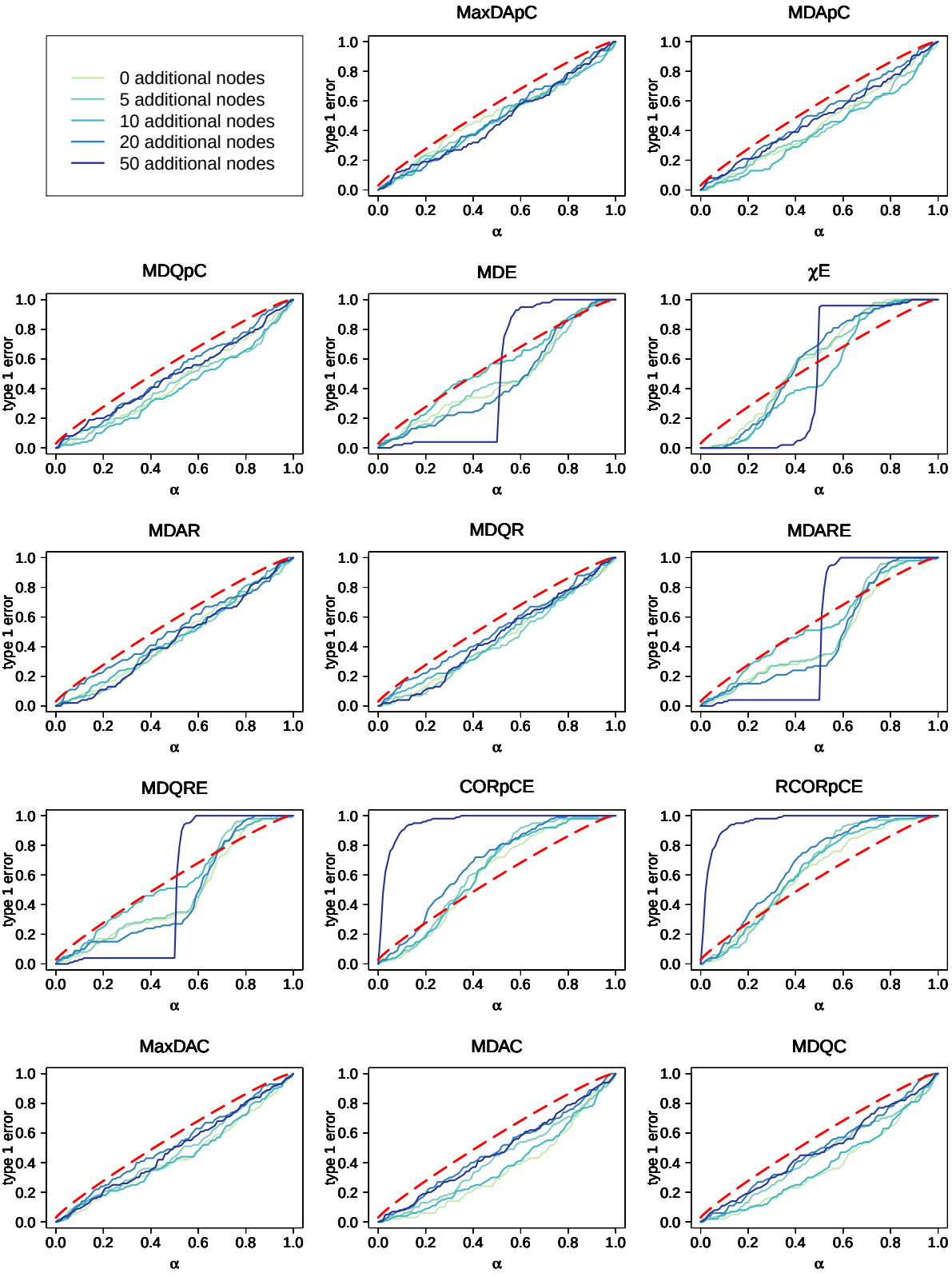

Figure 62: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 4.

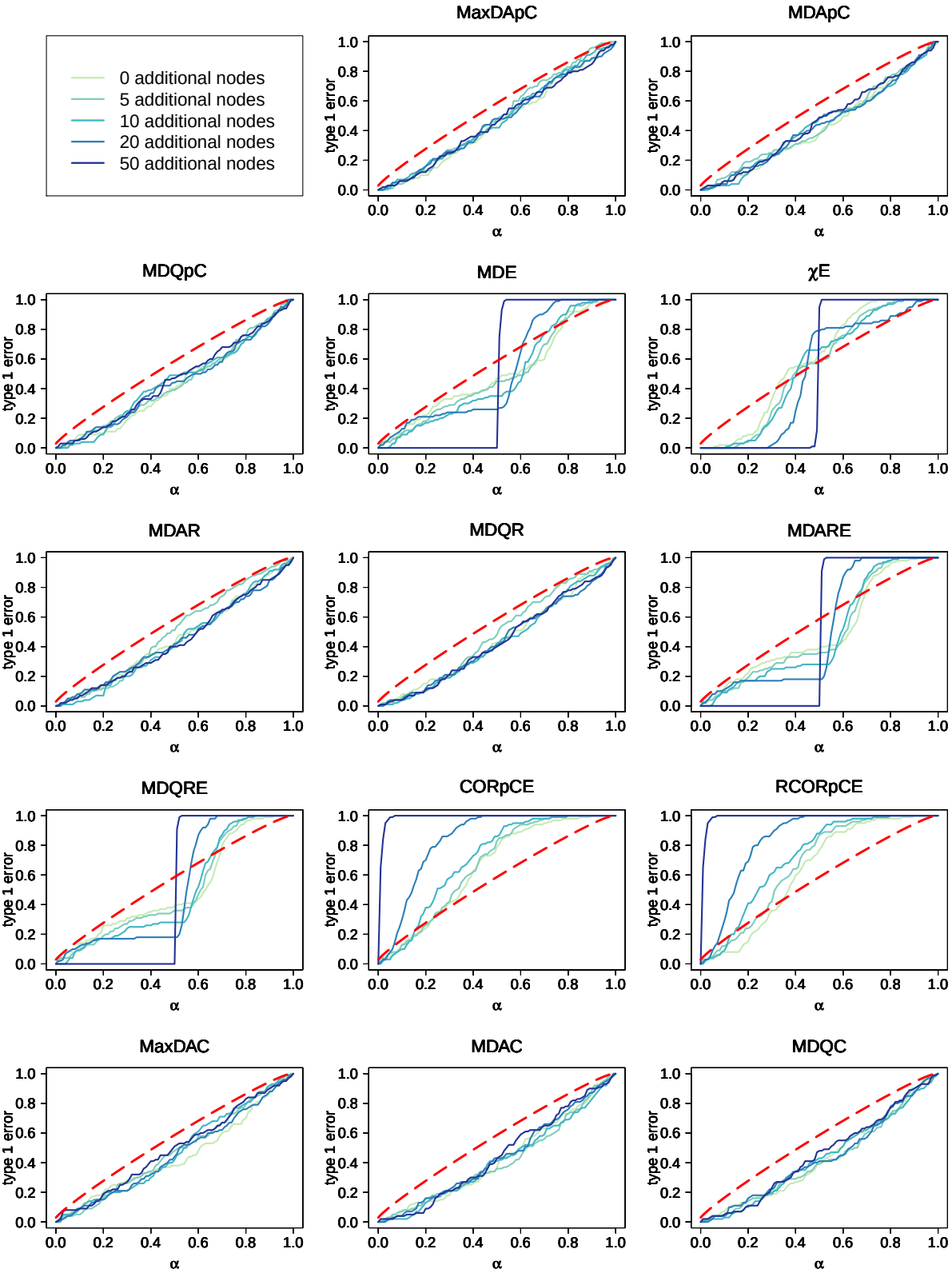

Figure 63: Proportion of misleadingly rejected hypothesis for simulated setting of 20 samples in each group and noise 8.

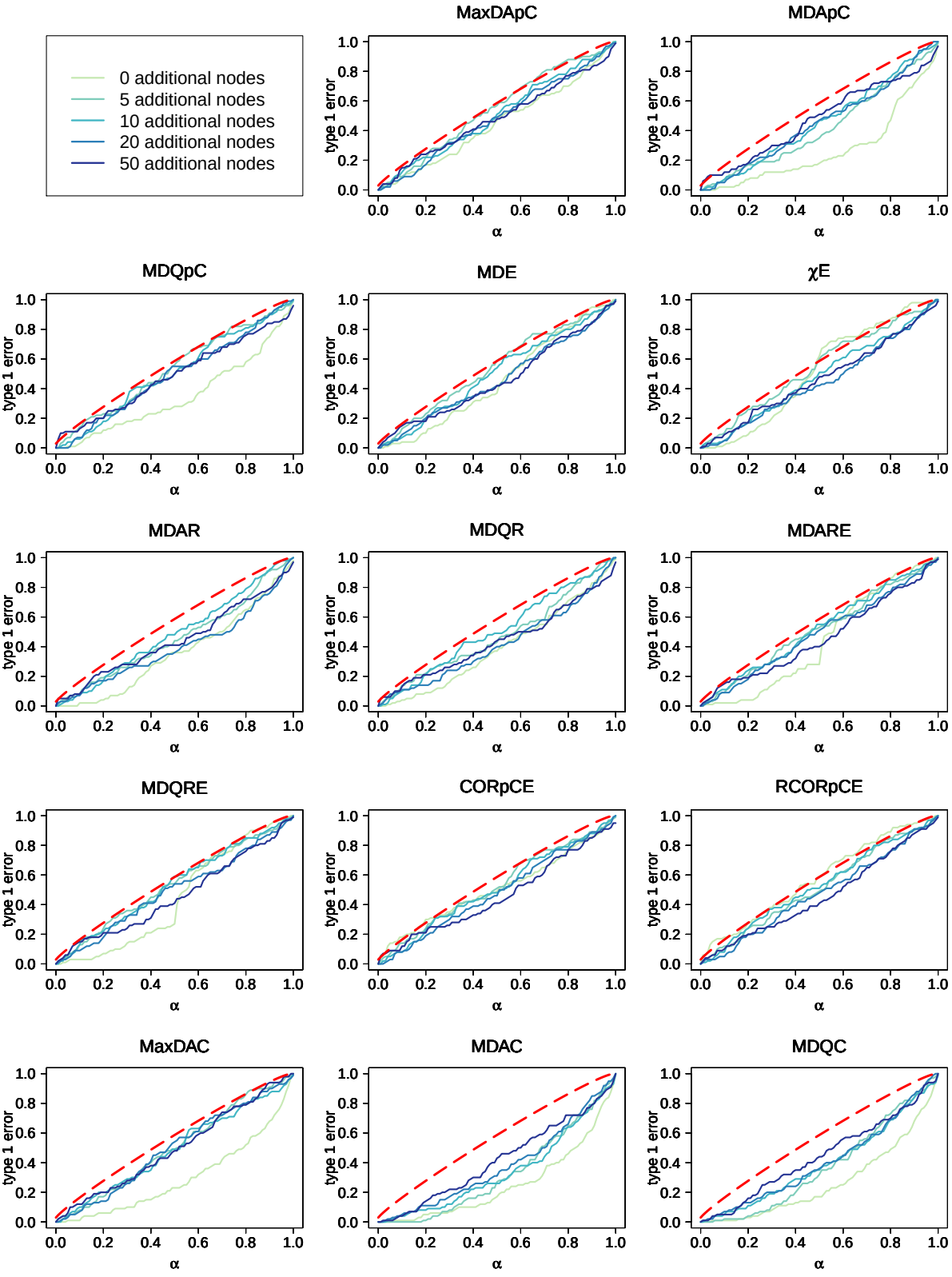

Figure 64: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 0.01.

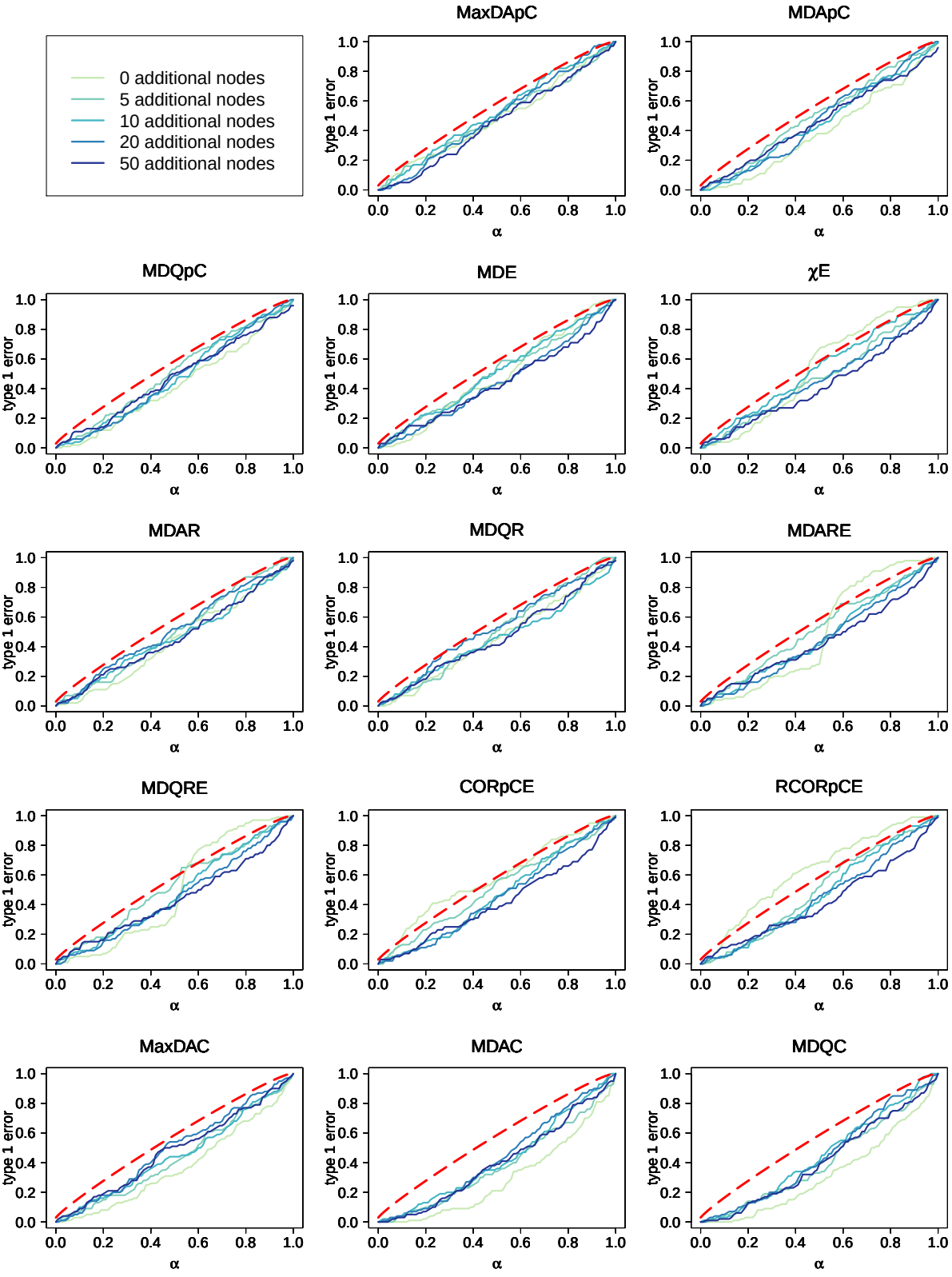

Figure 65: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 0.1.

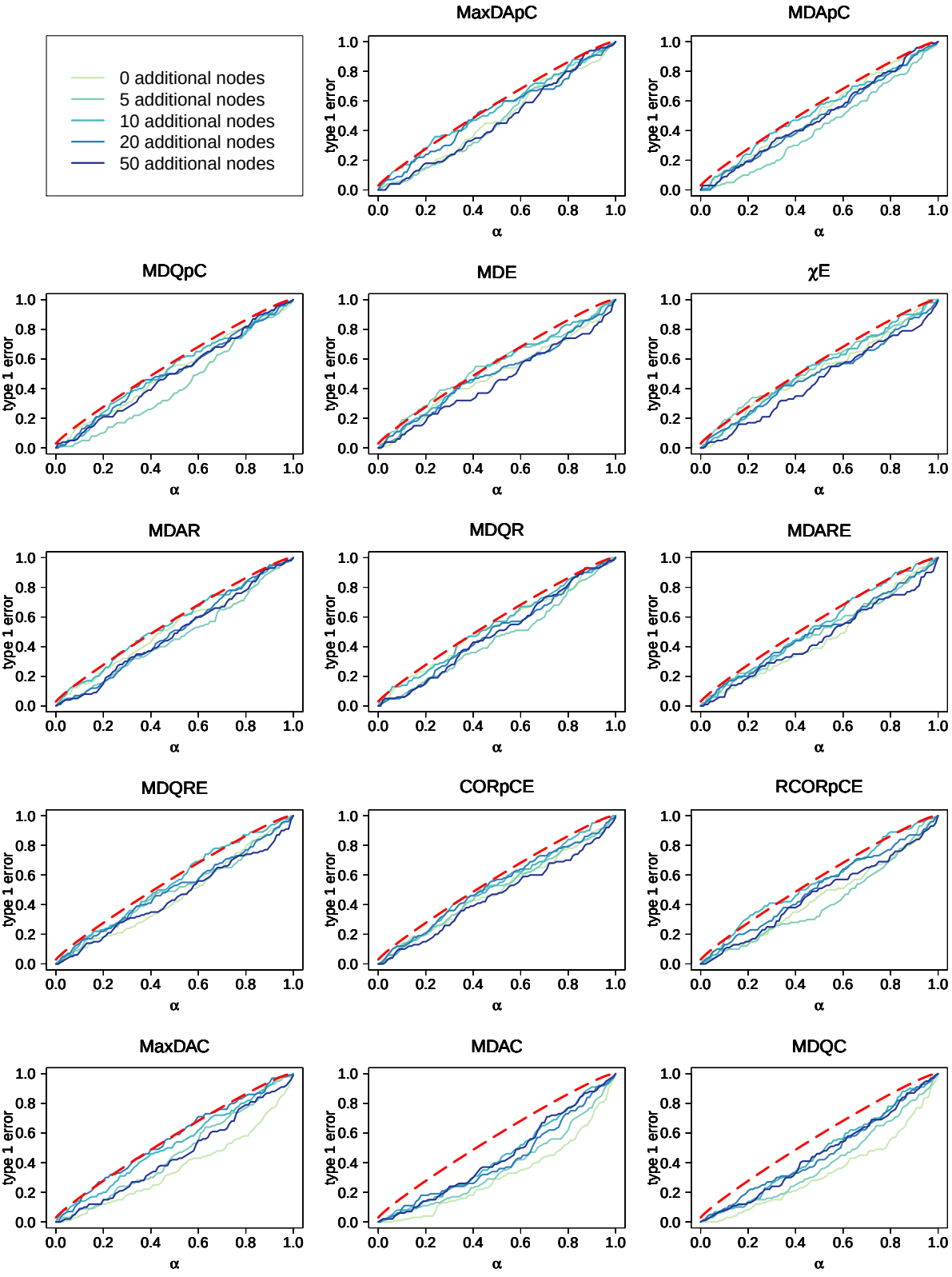

Figure 66: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 0.5.

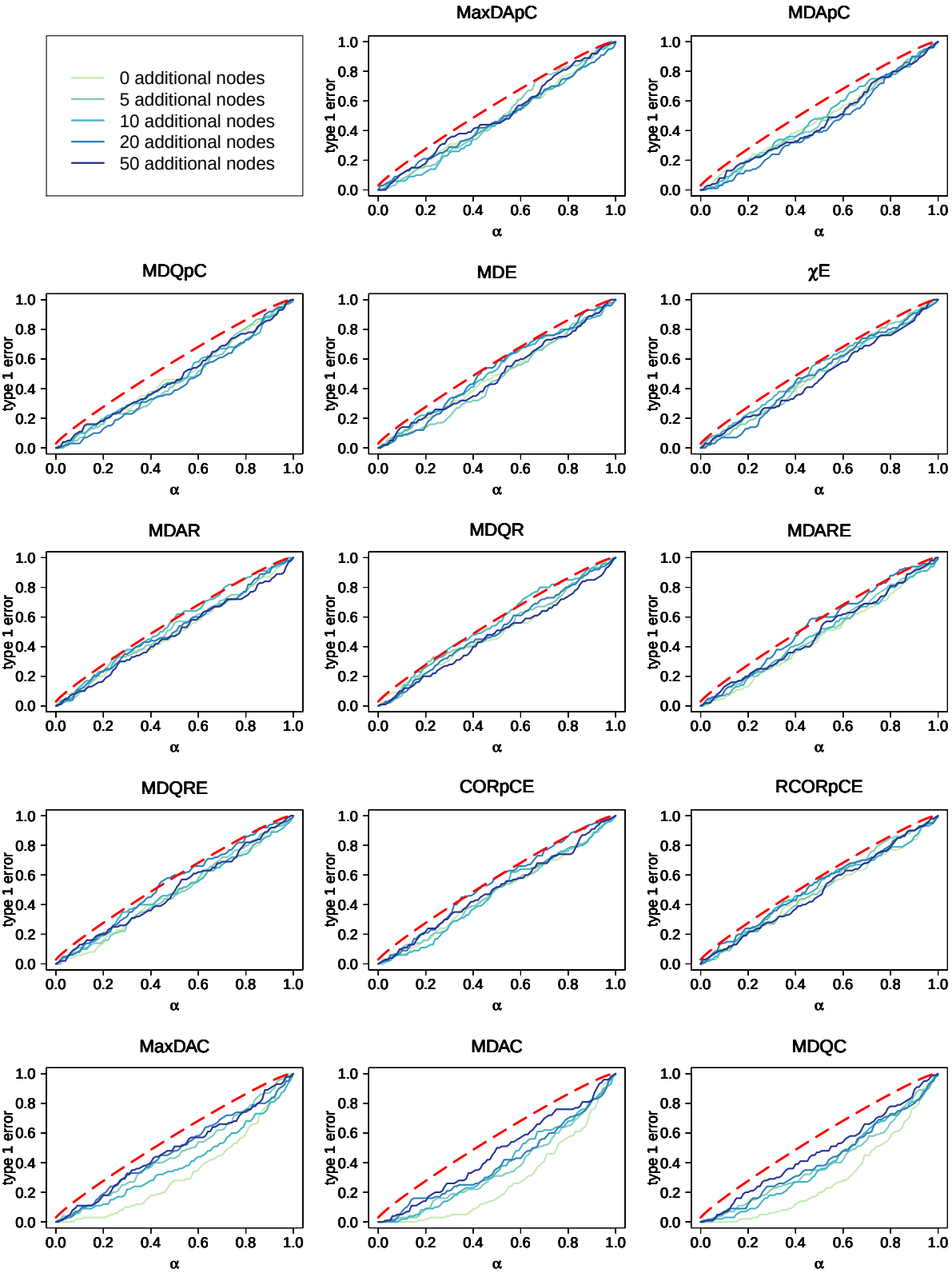

Figure 67: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 1.

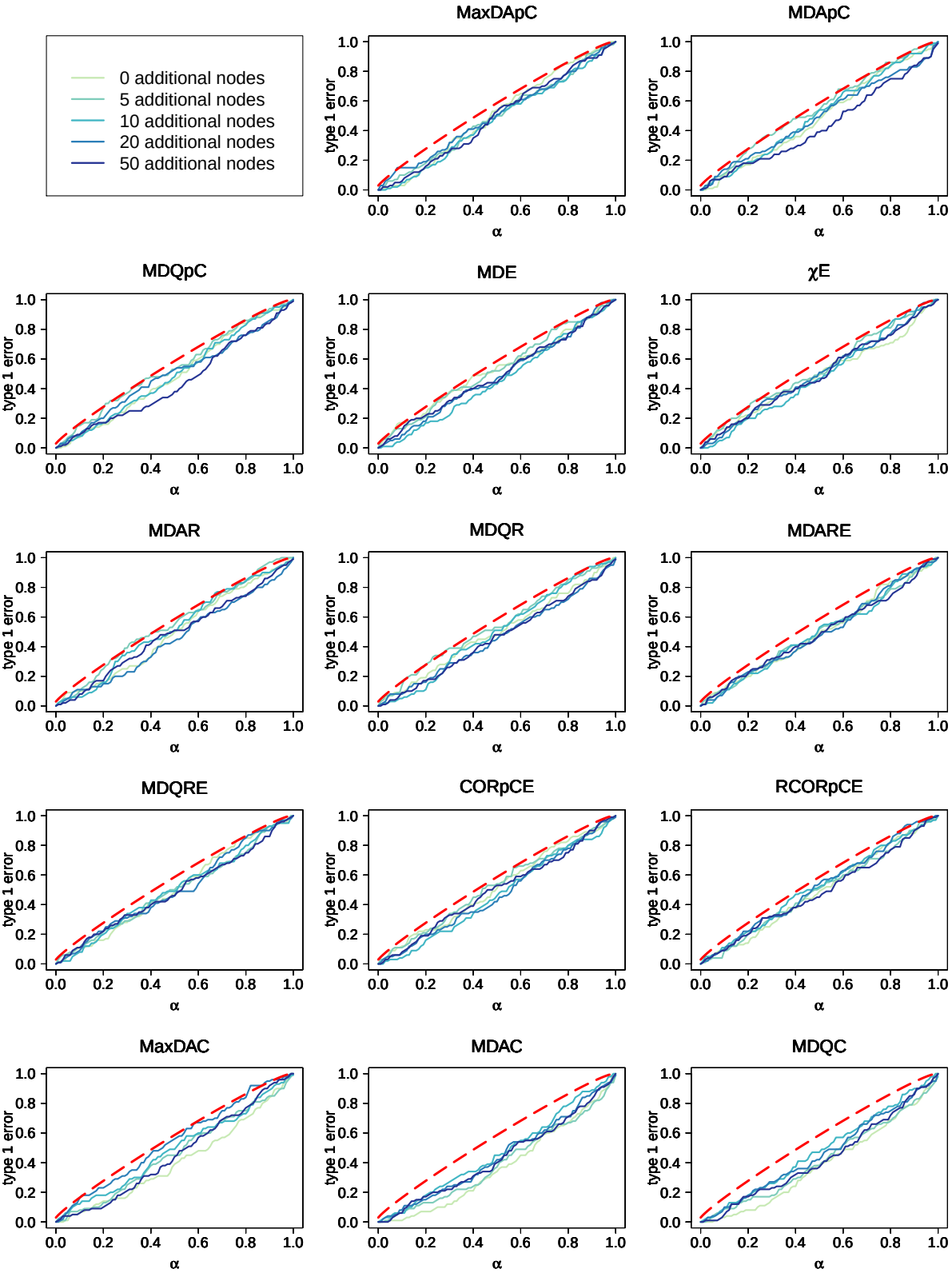

Figure 68: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 2.

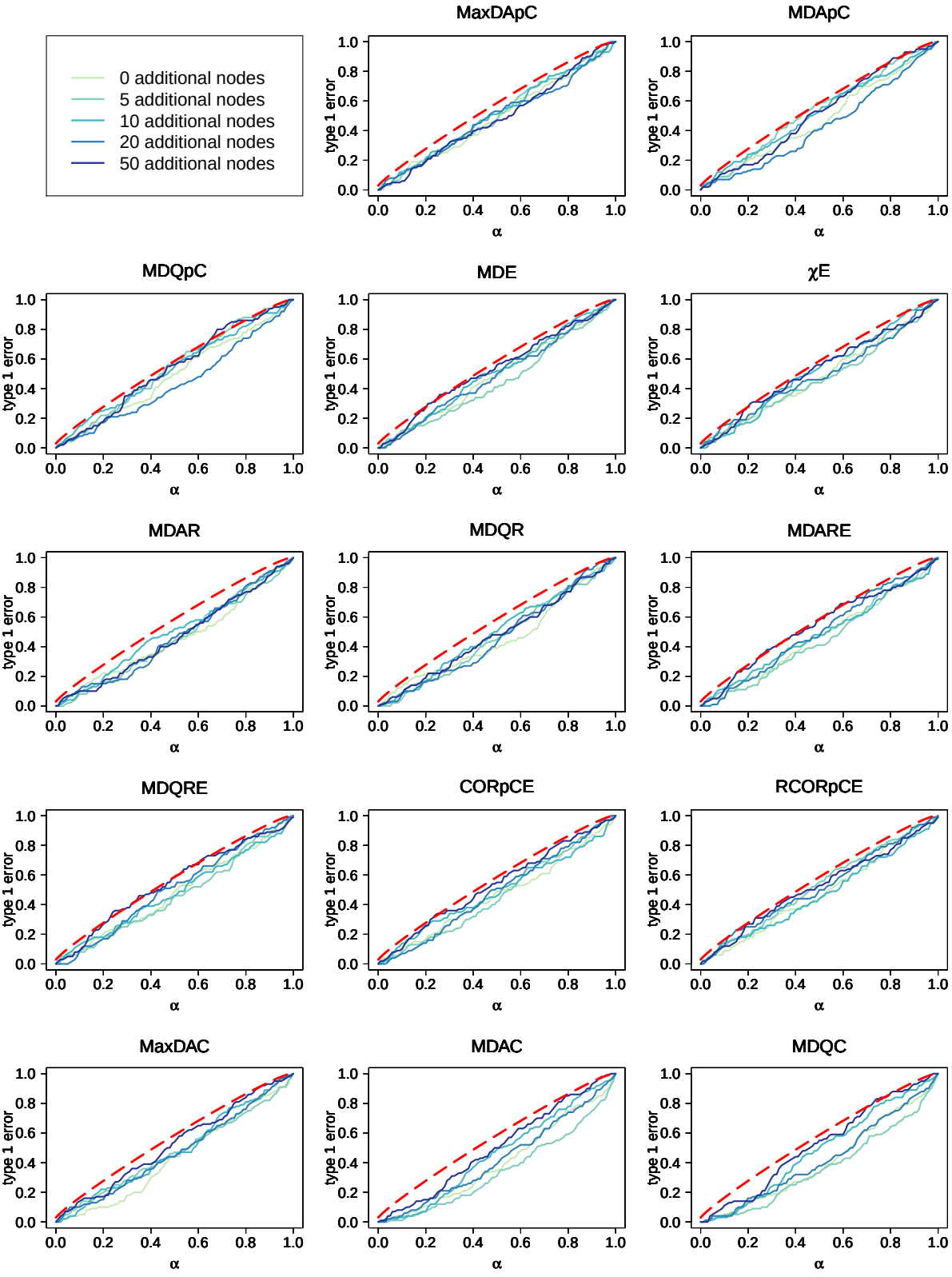

Figure 69: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 4.

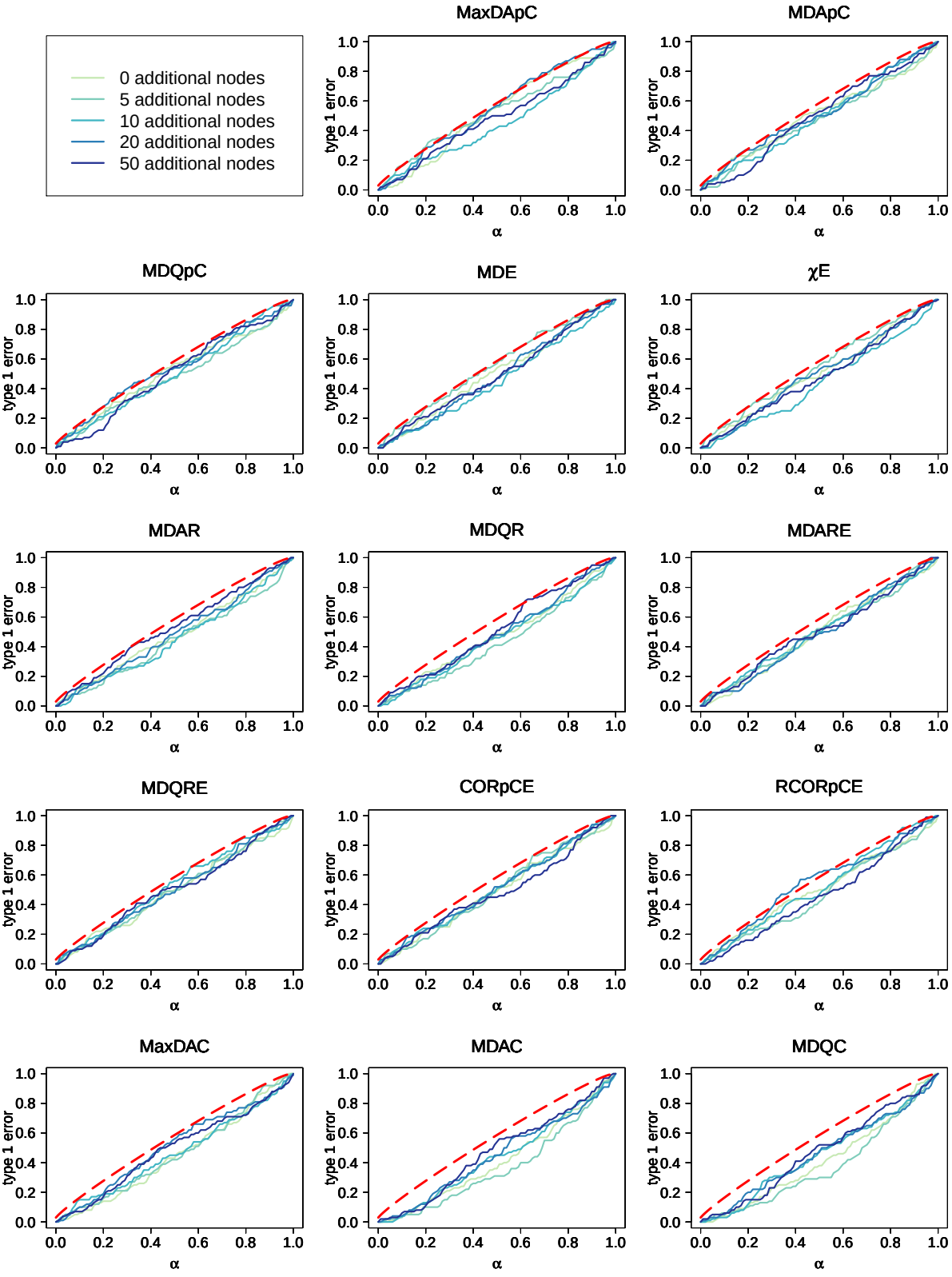

Figure 70: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 8.

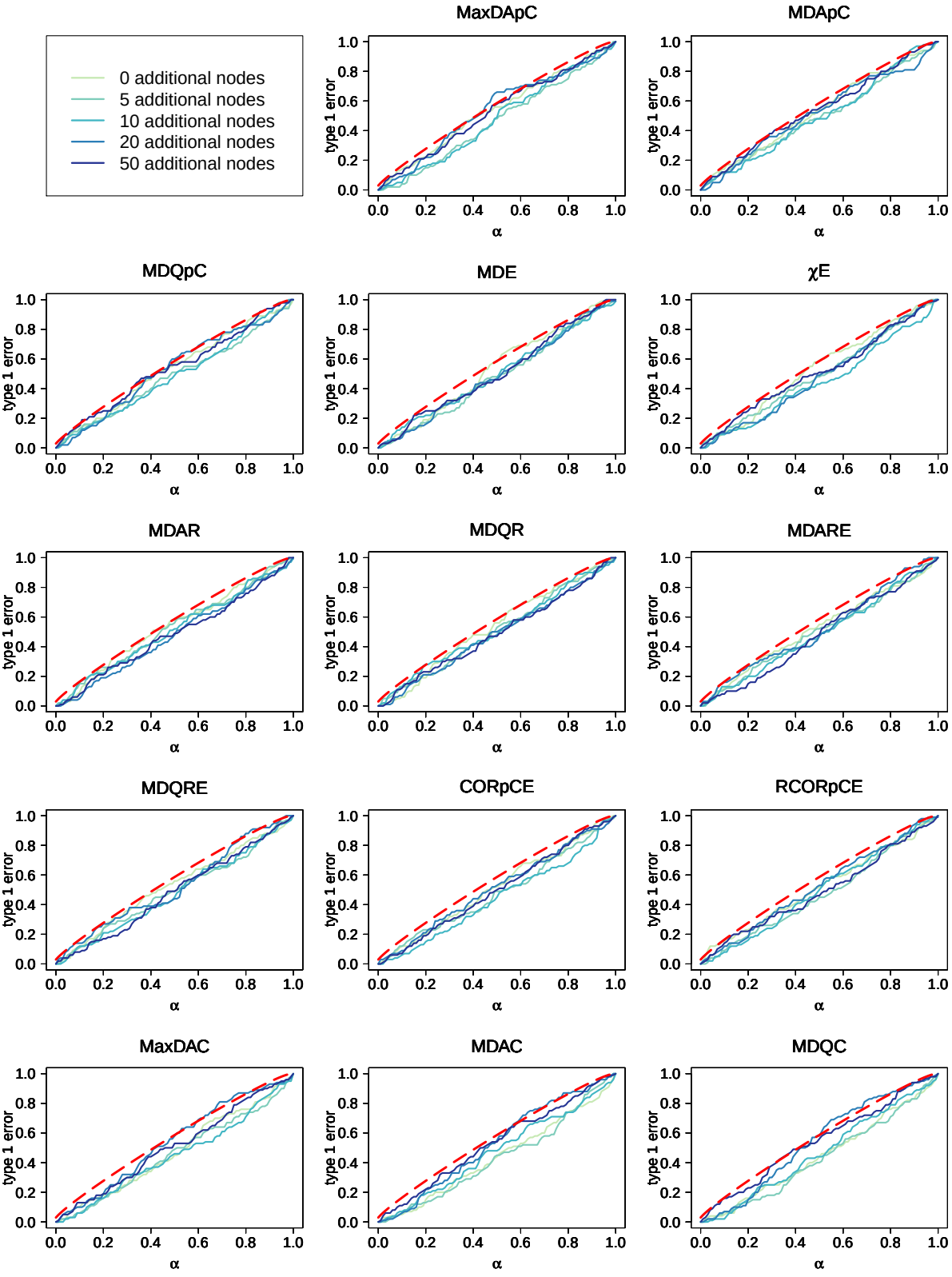

Figure 71: Proportion of misleadingly rejected hypothesis for simulated setting of 200 samples in each group and noise 16.

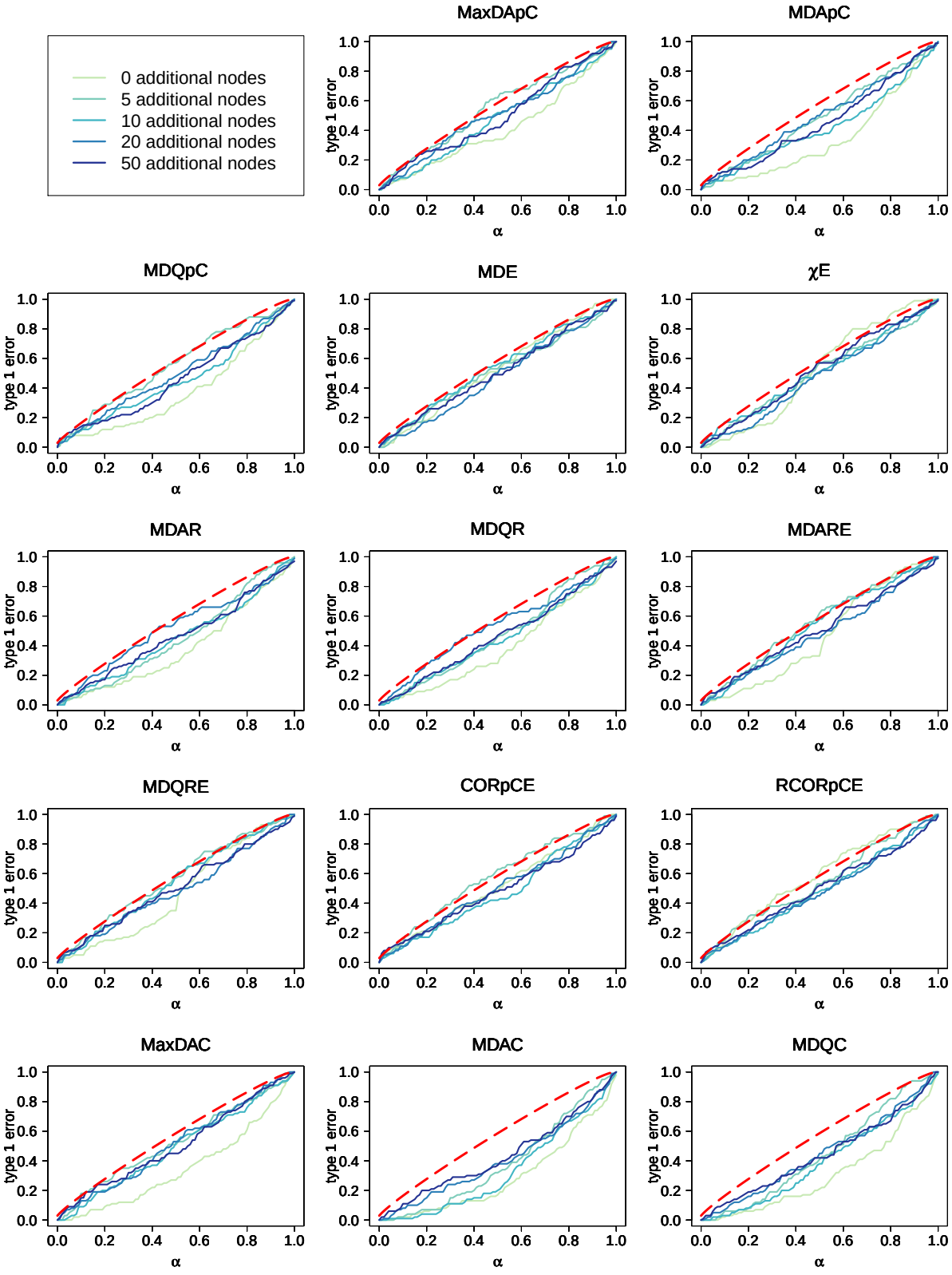

Figure 72: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 0.01.

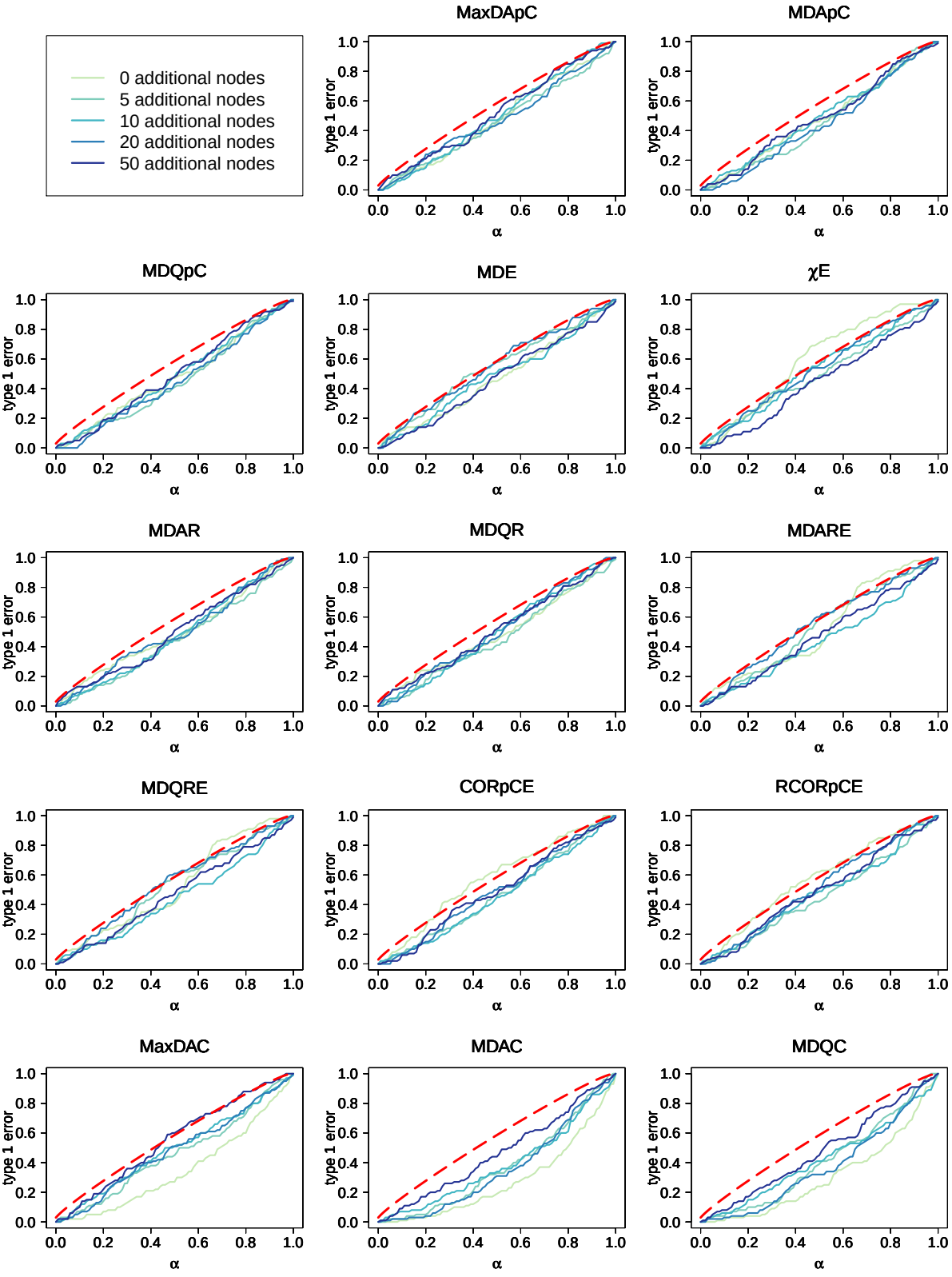

Figure 73: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 0.1.

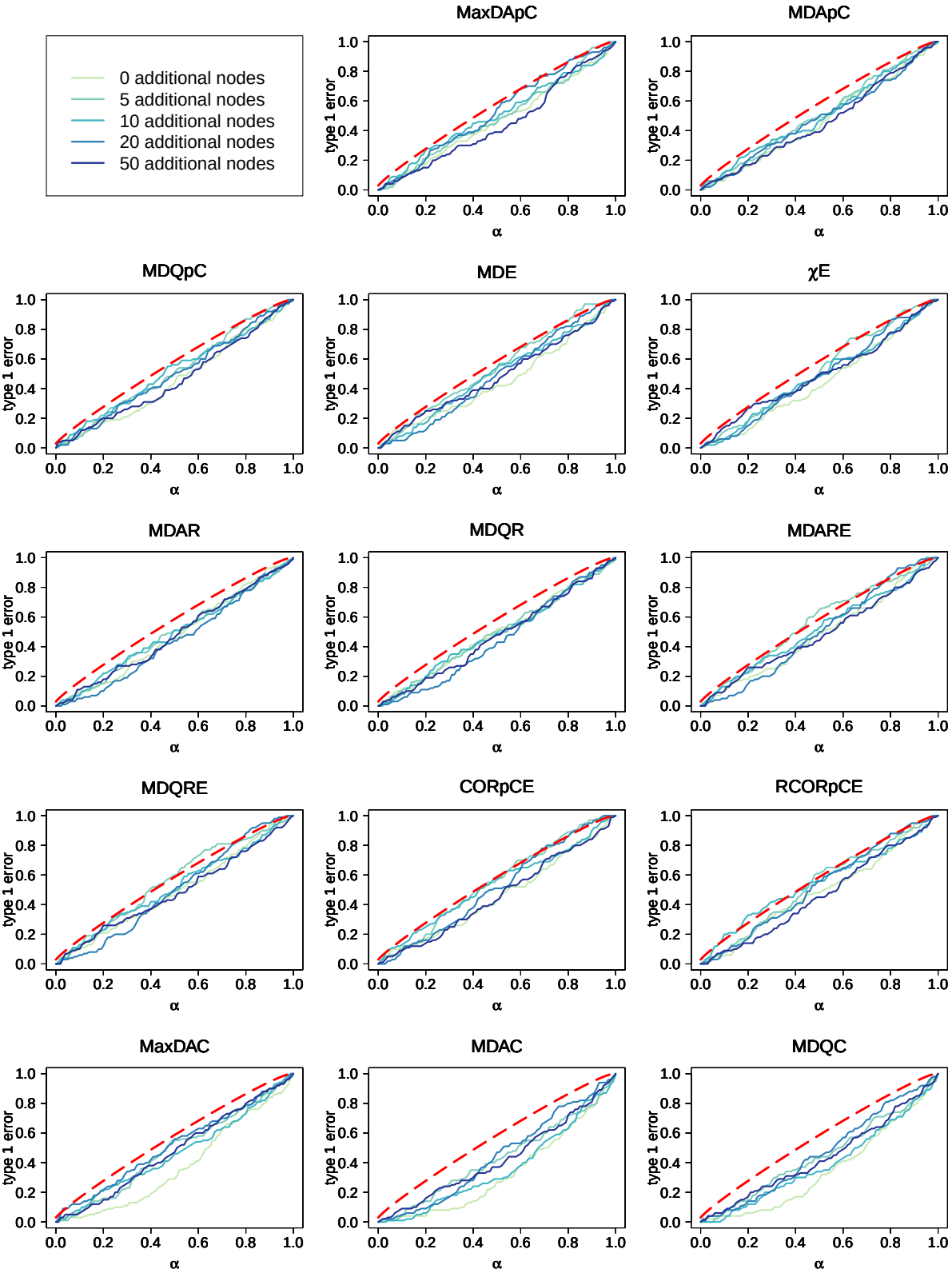

Figure 74: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 0.5.

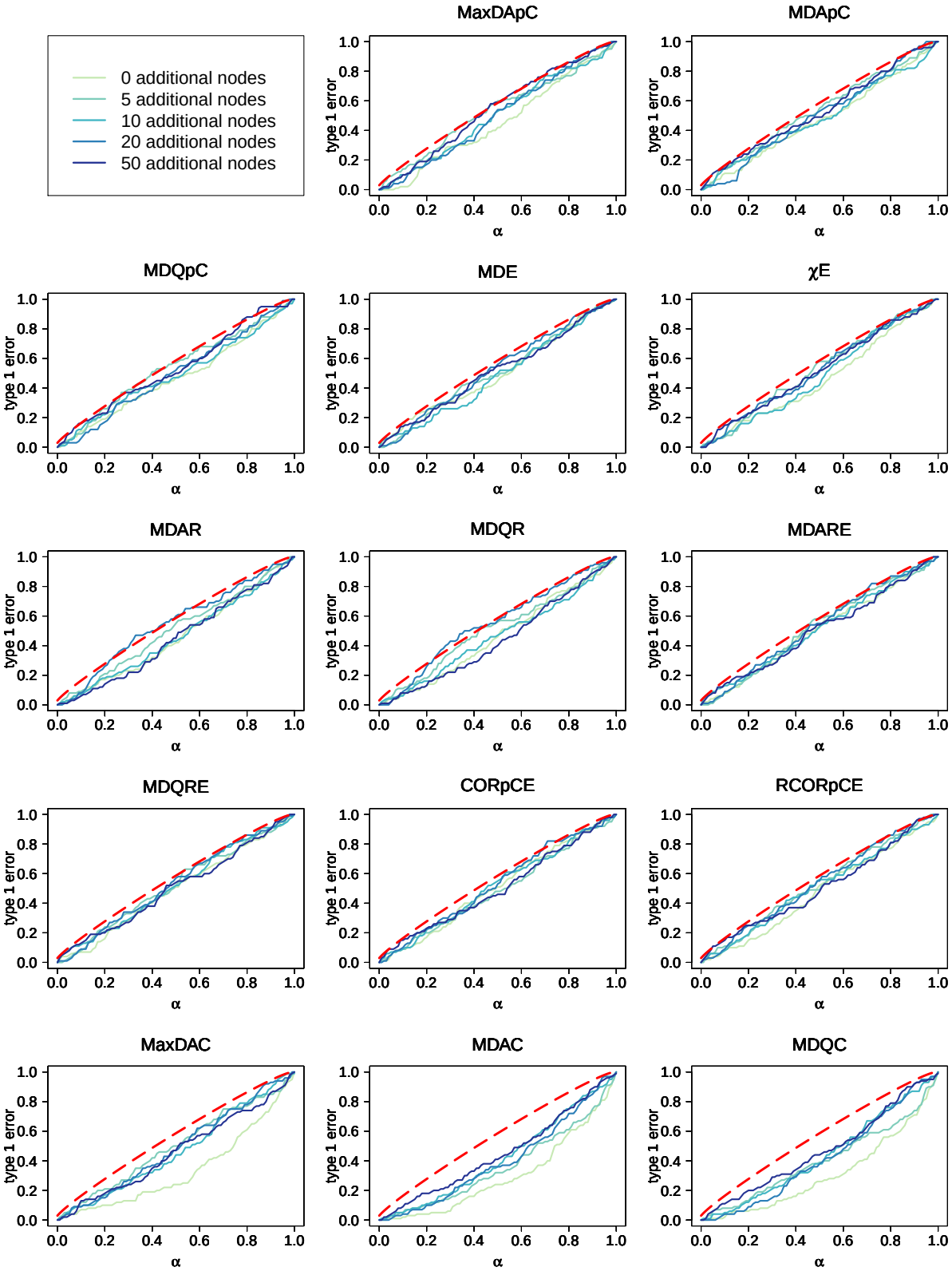

Figure 75: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 1.

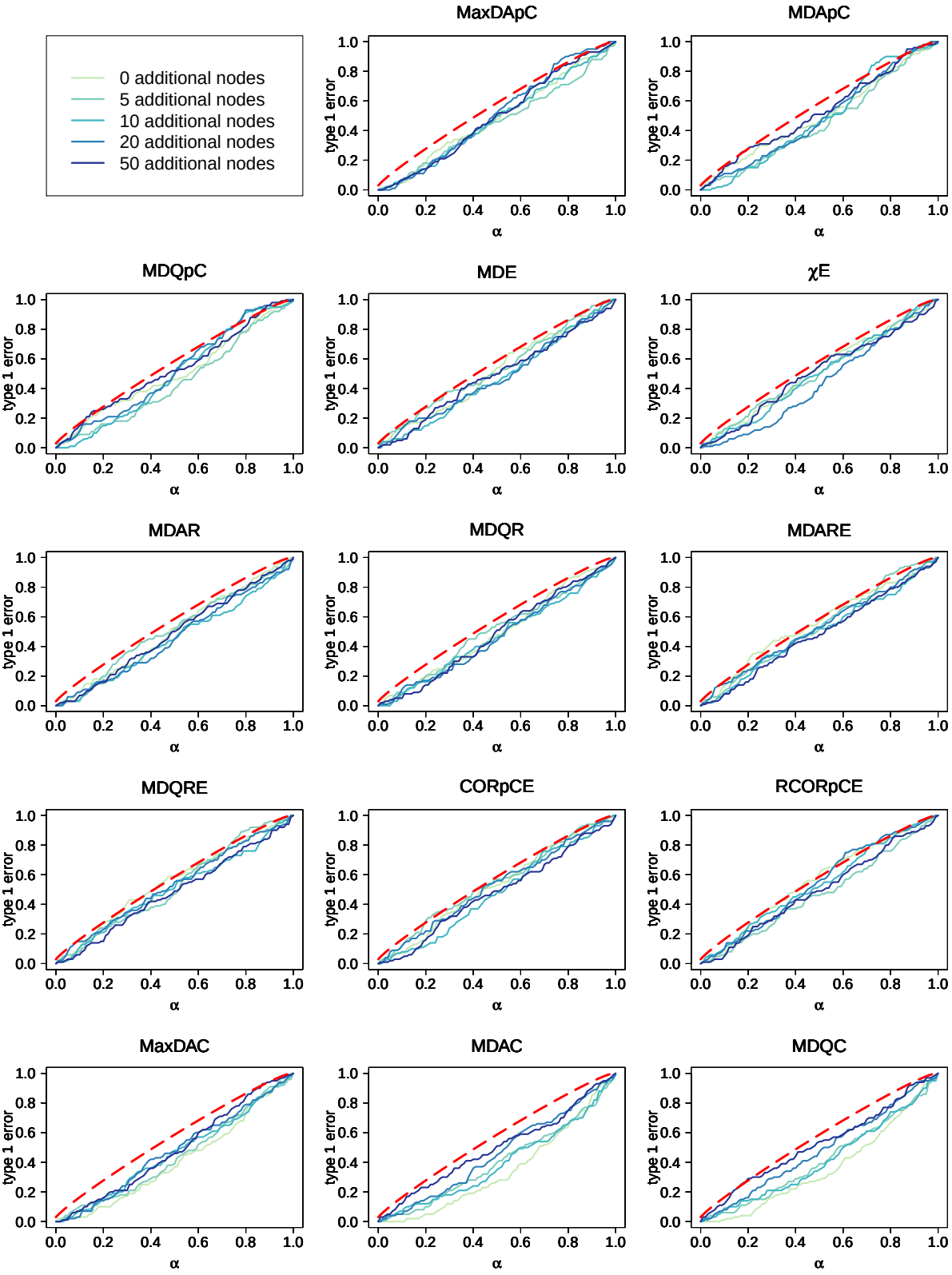

Figure 76: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 2.

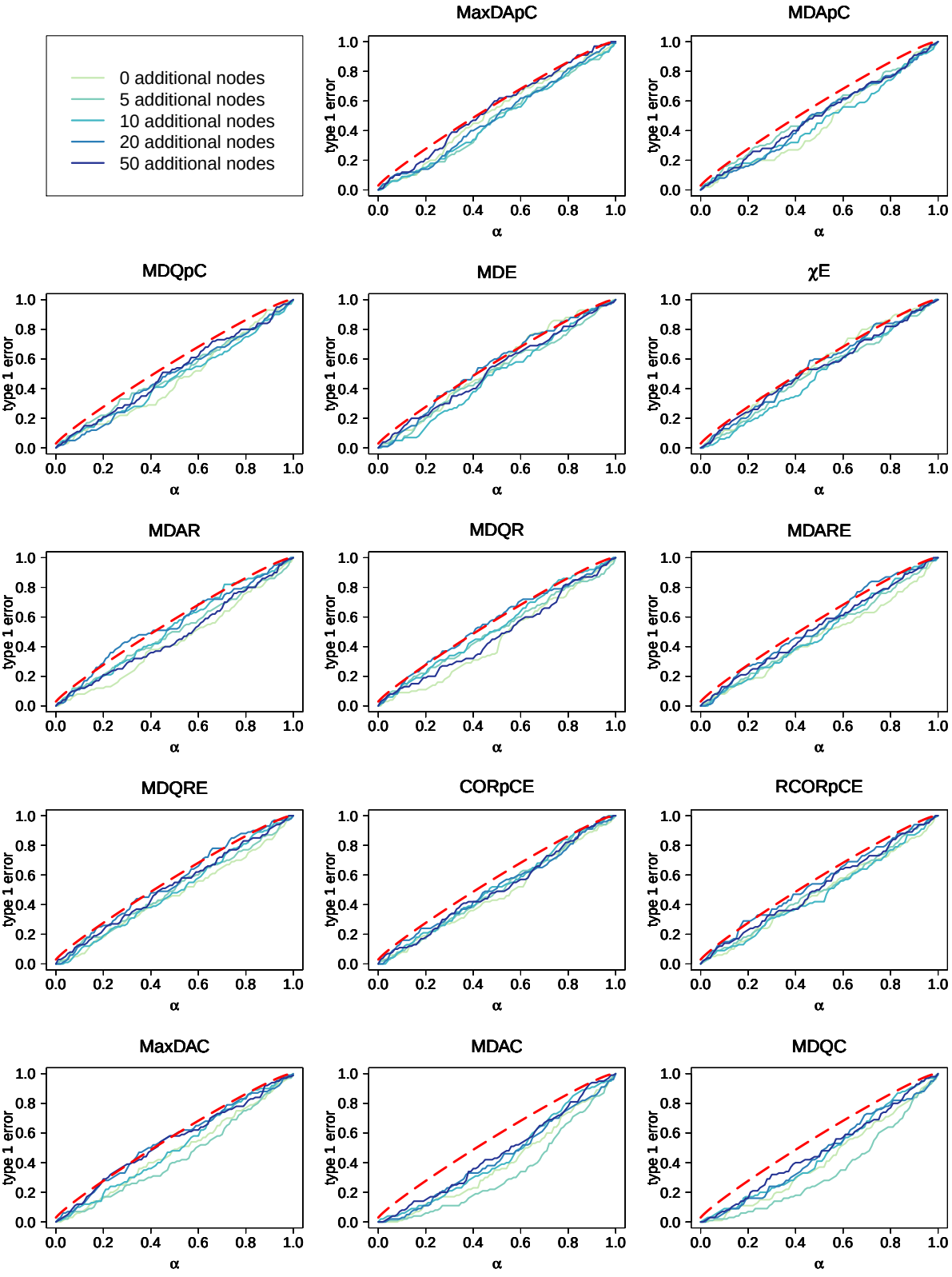

Figure 77: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 4.

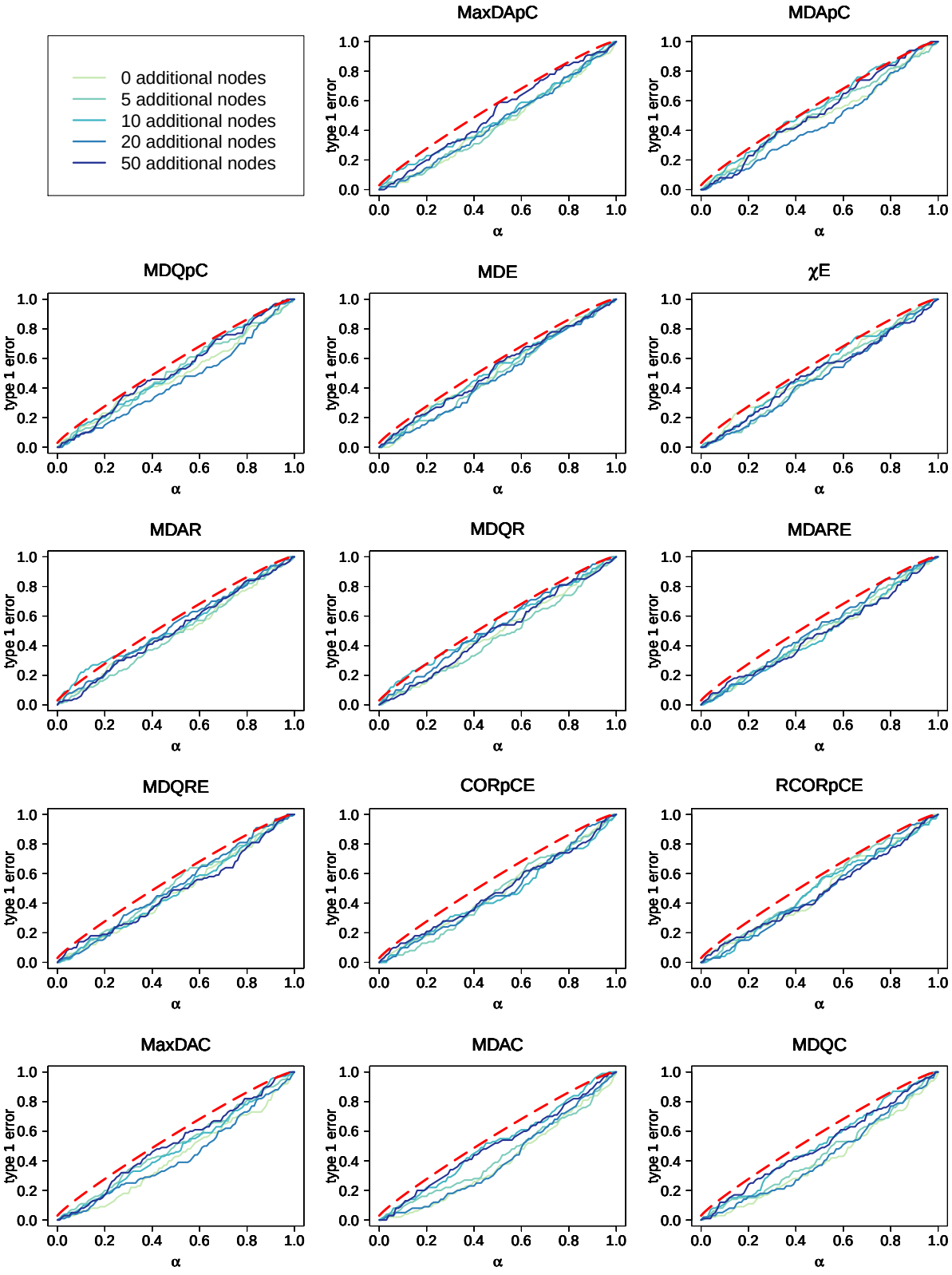

Figure 78: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 4.

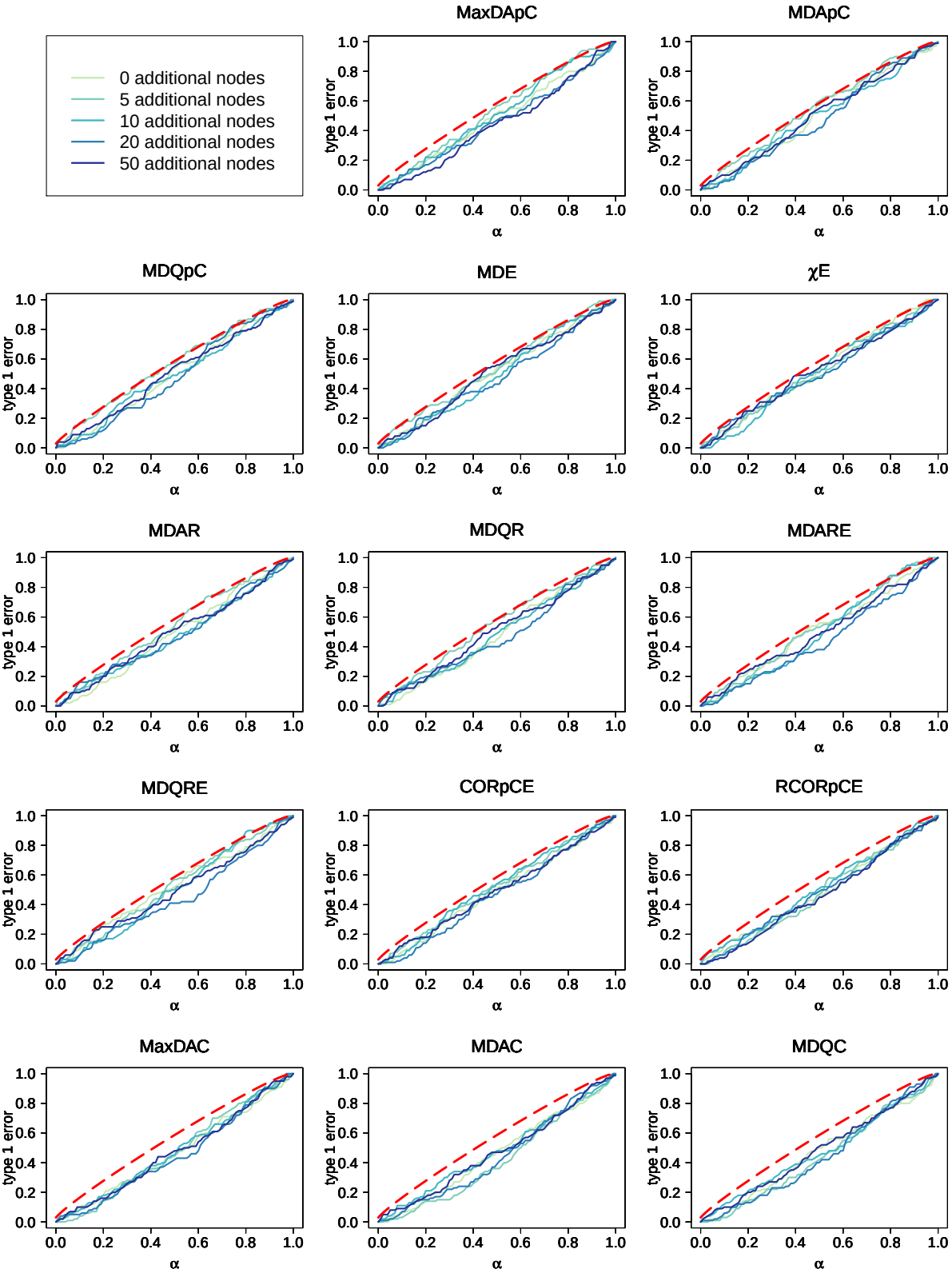

Figure 79: Proportion of misleadingly rejected hypothesis for simulated setting of 300 and 100 samples per group and noise 16.

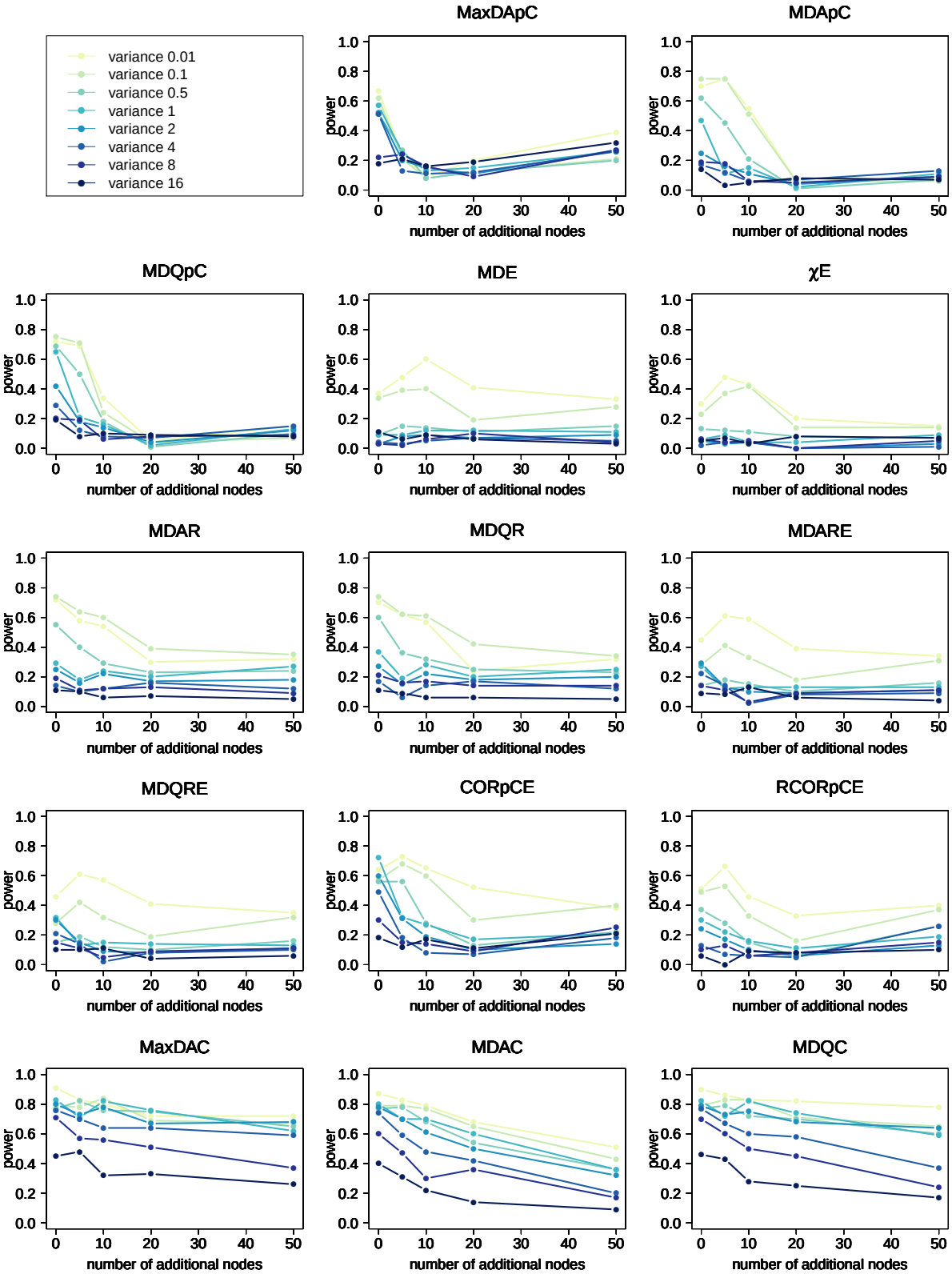

Figure 80: Proportion of rejected hypothesis for simulated setting of 150 and 50 samples and knockout of node PKC.

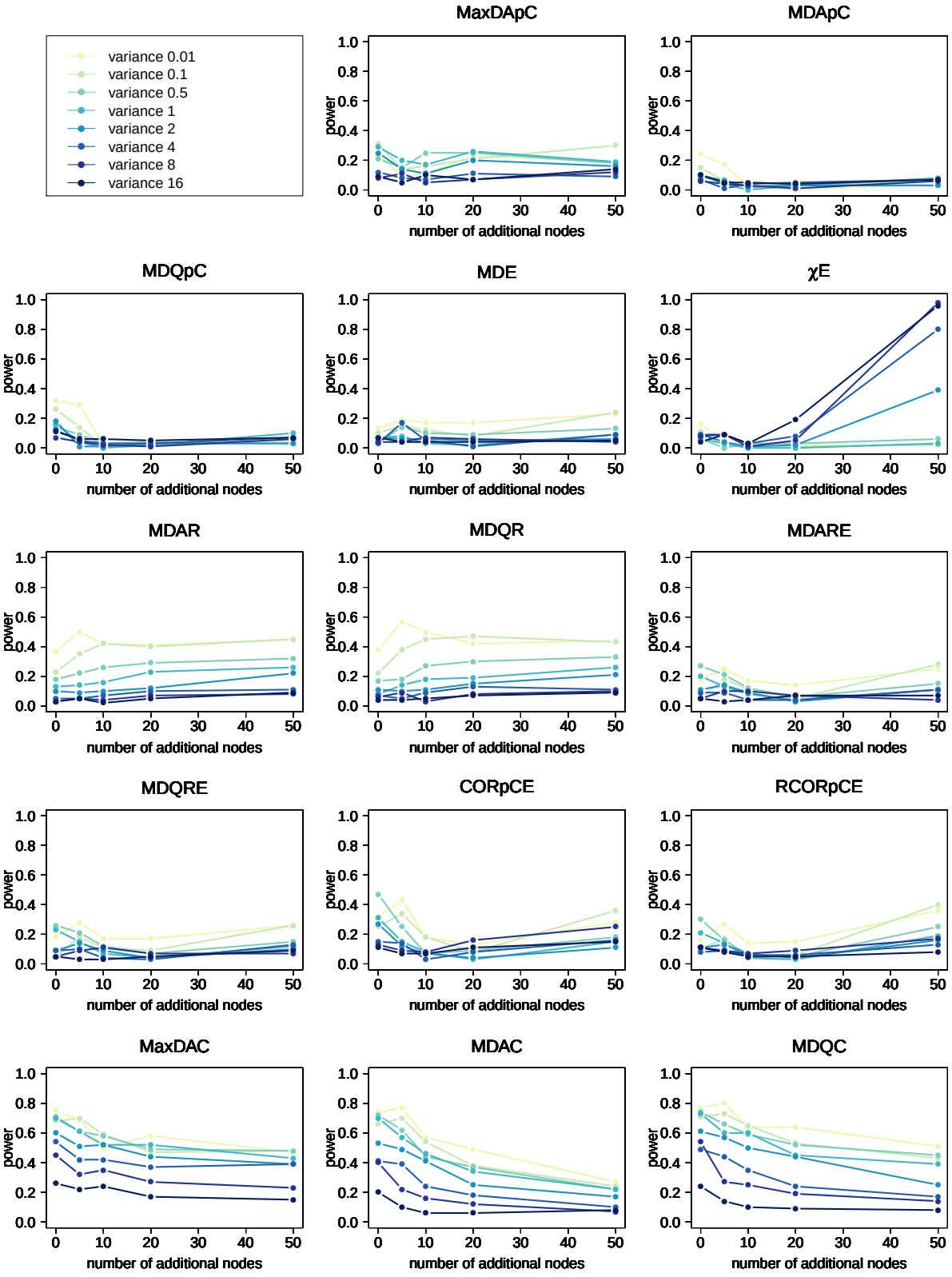

Figure 81: Proportion of rejected hypothesis for simulated setting of 180 and 20 samples and knockout of node PKC.

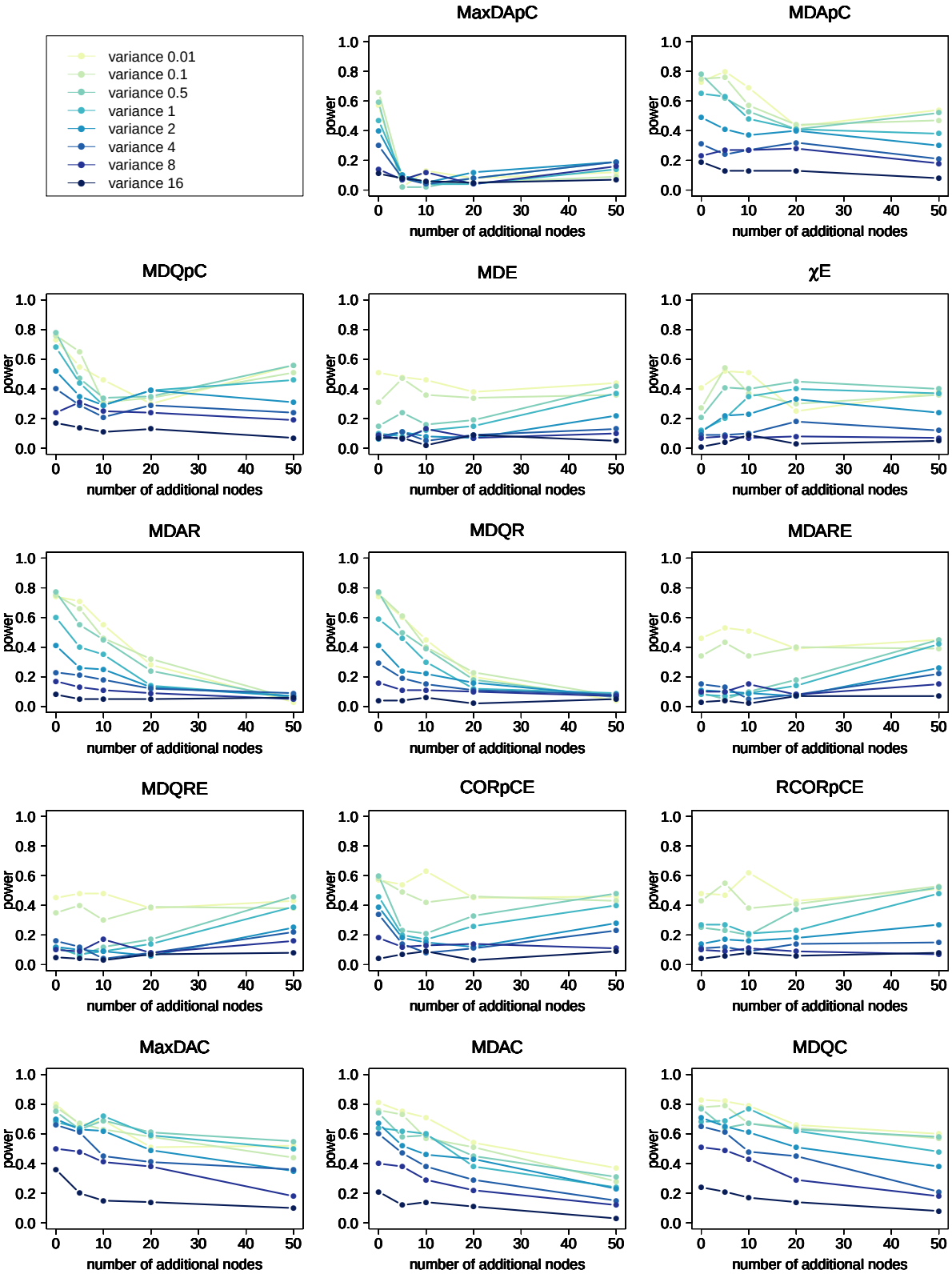

Figure 82: Proportion of rejected hypothesis for simulated setting of 50 samples per group and knockout of node PKC.

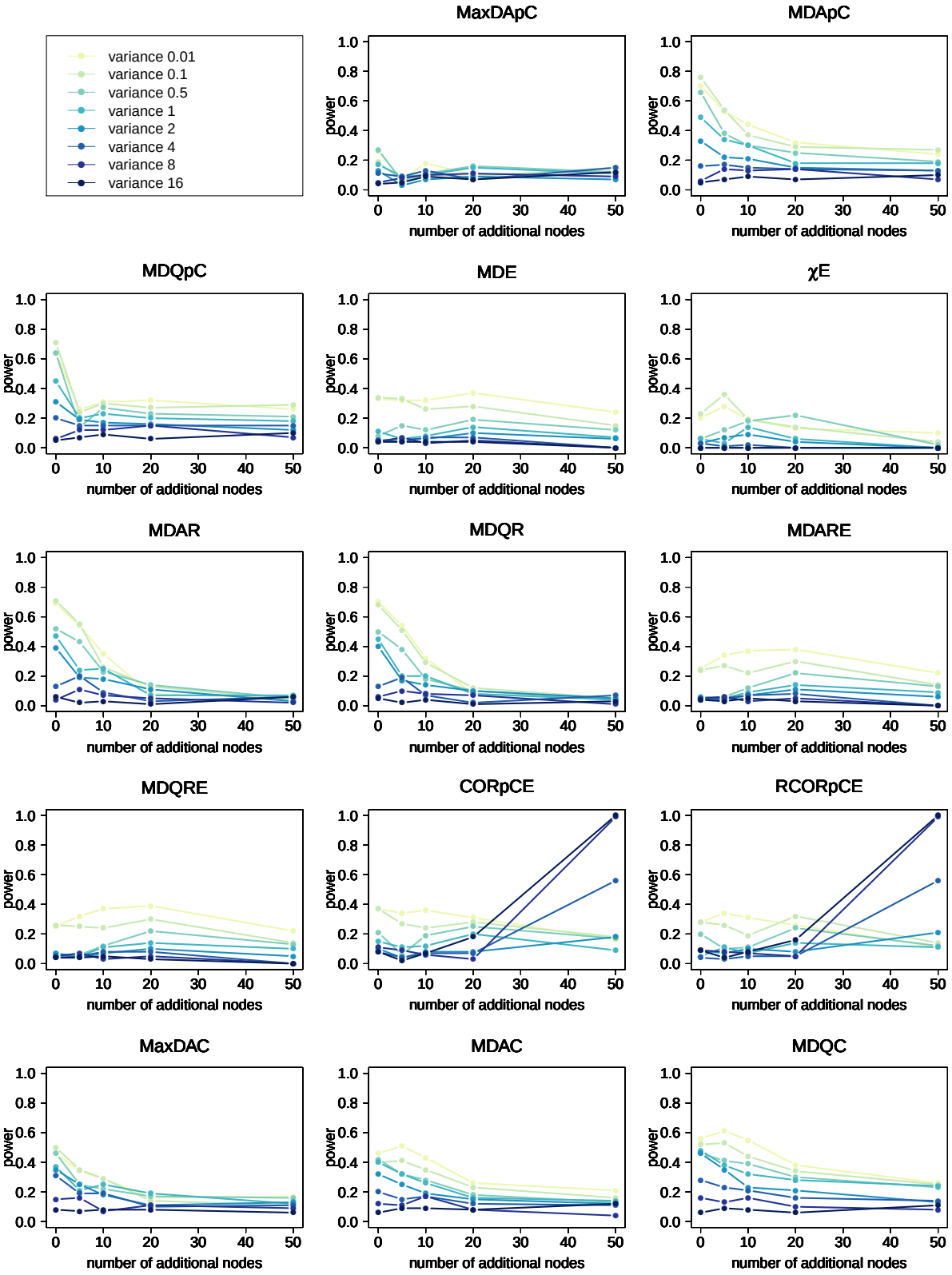

Figure 83: Proportion of rejected hypothesis for simulated setting of 20 samples per group and knockout of node PKC.

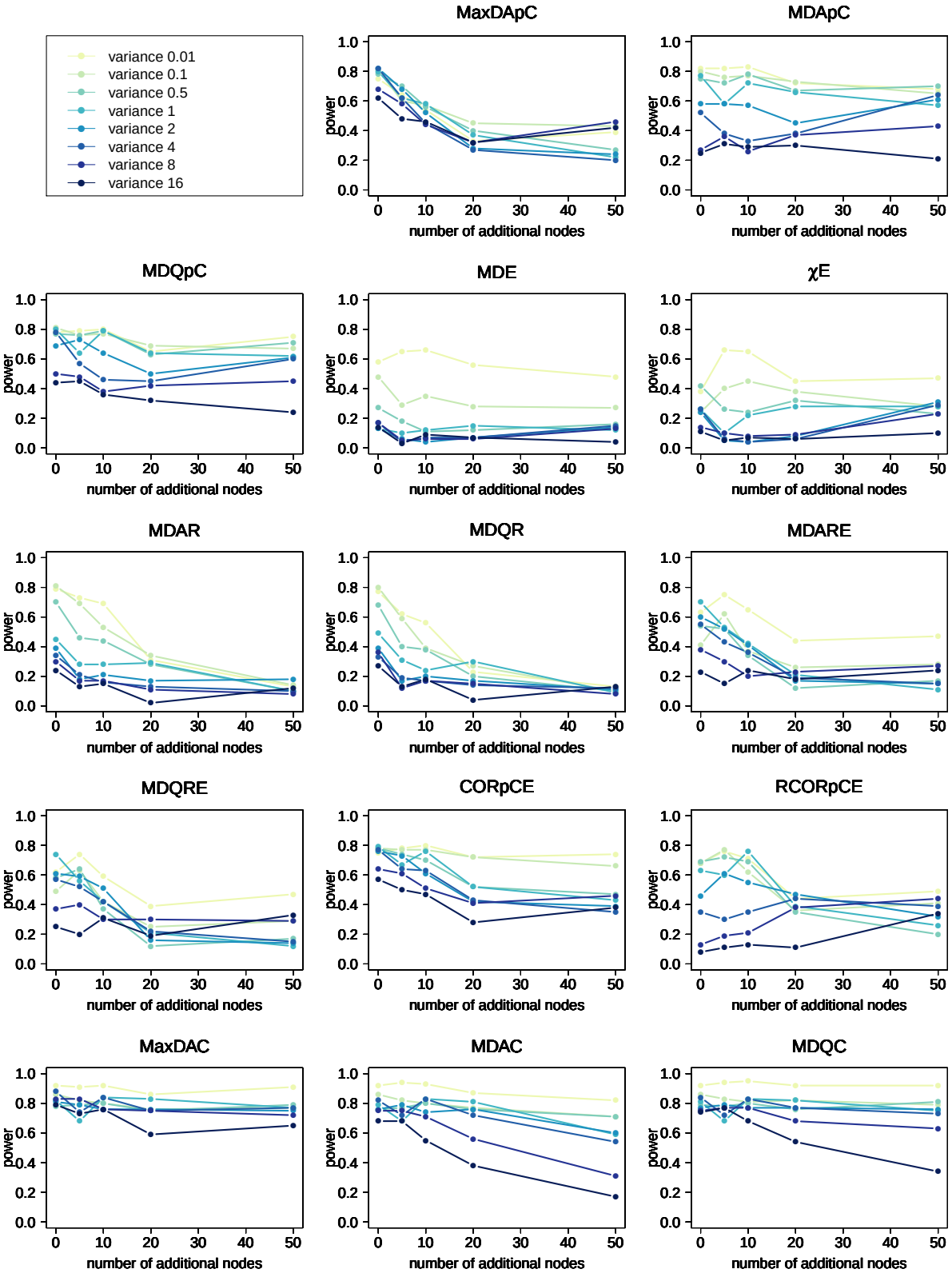

Figure 84: Proportion of rejected hypothesis for simulated setting of 200 samples per group and knockout of node PKC.

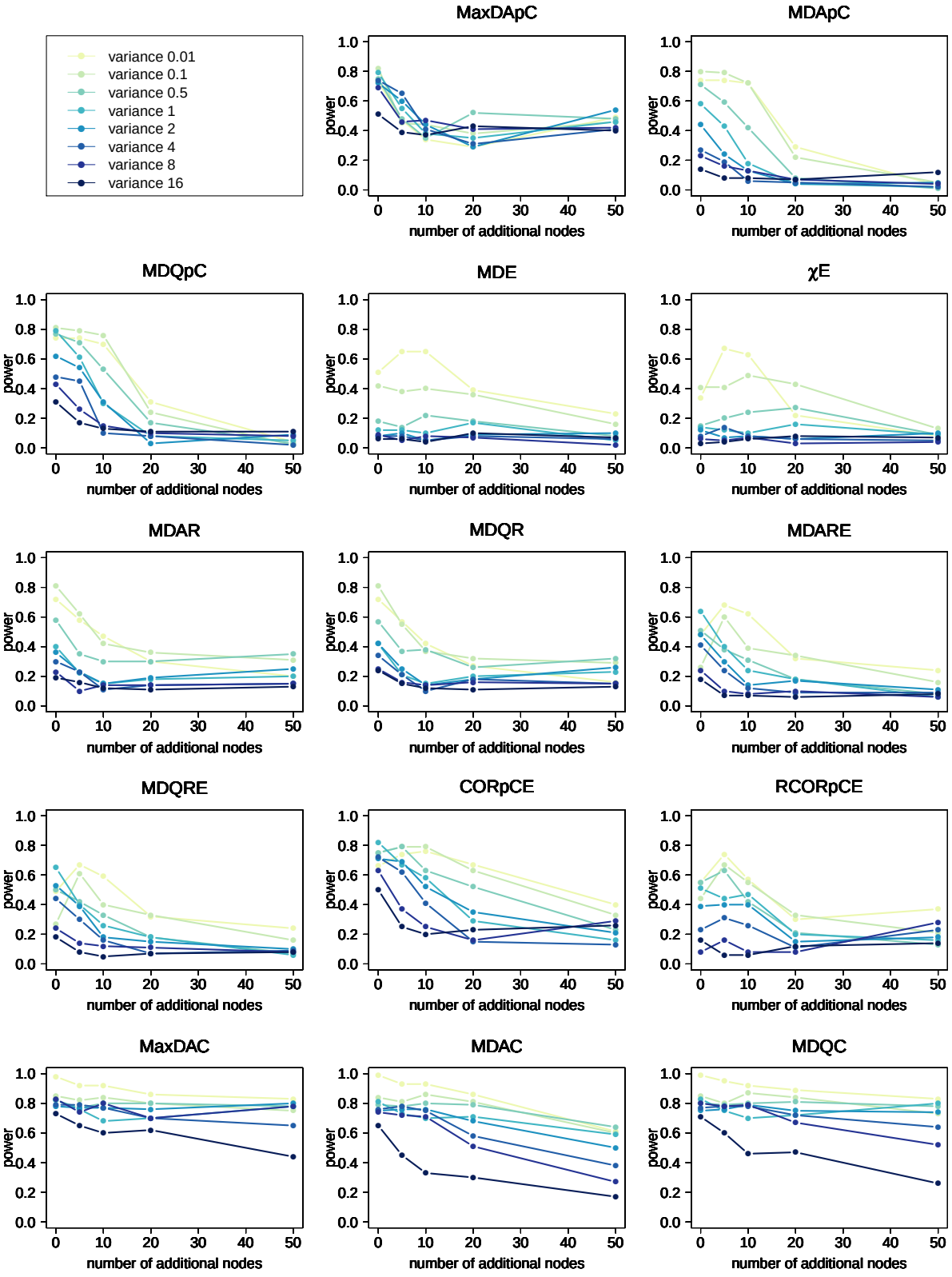

Figure 85: Proportion of rejected hypothesis for simulated setting of 300 and 100 samples and knockout of node PKC.

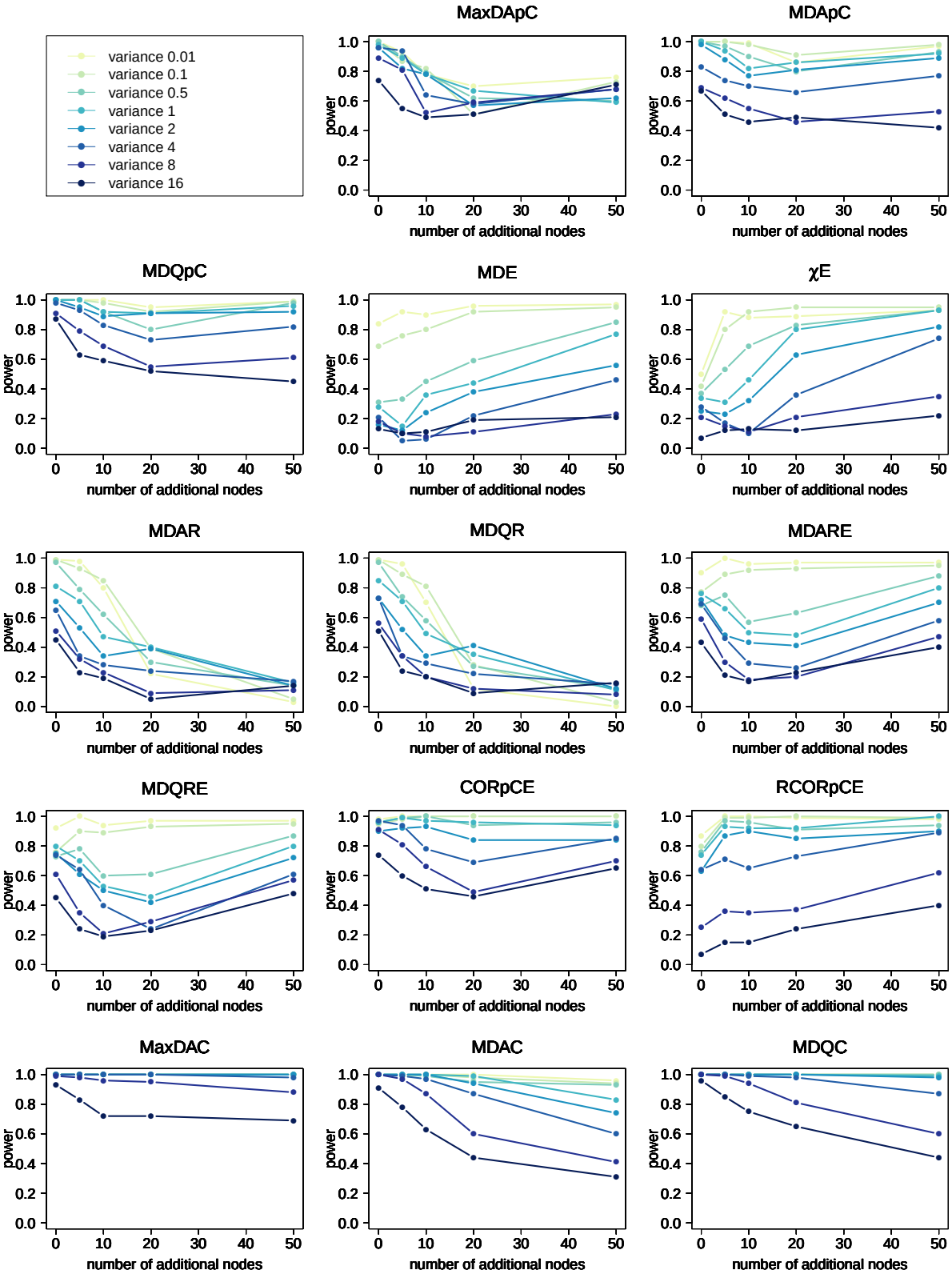

Figure 86: Proportion of rejected hypothesis for simulated setting of 100 samples per group and knockout of nodes PIP2, PKC, and PKA.

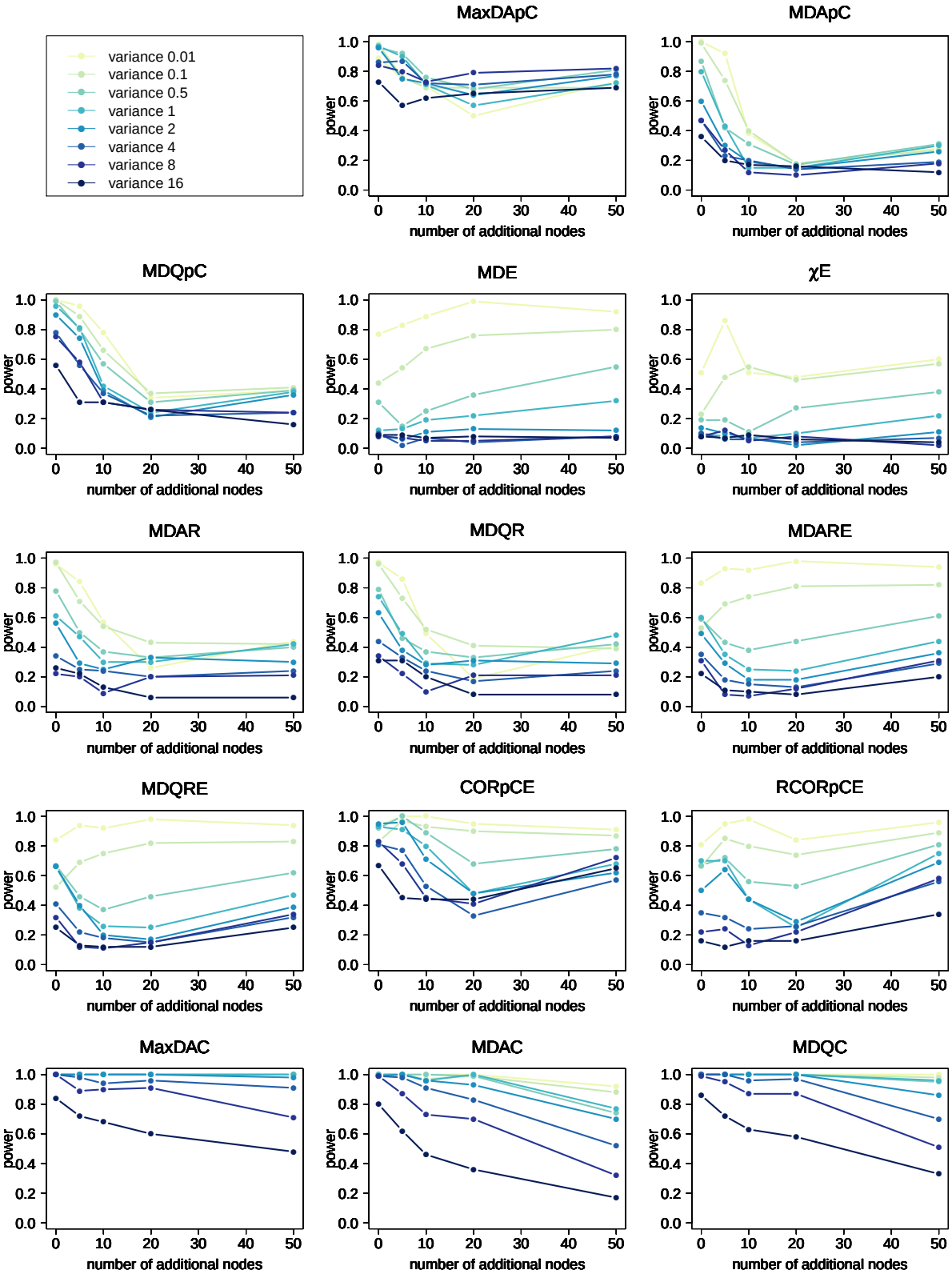

Figure 87: Proportion of rejected hypothesis for simulated setting of 150 and 50 samples and knockout of nodes PIP2, PKC, and PKA.

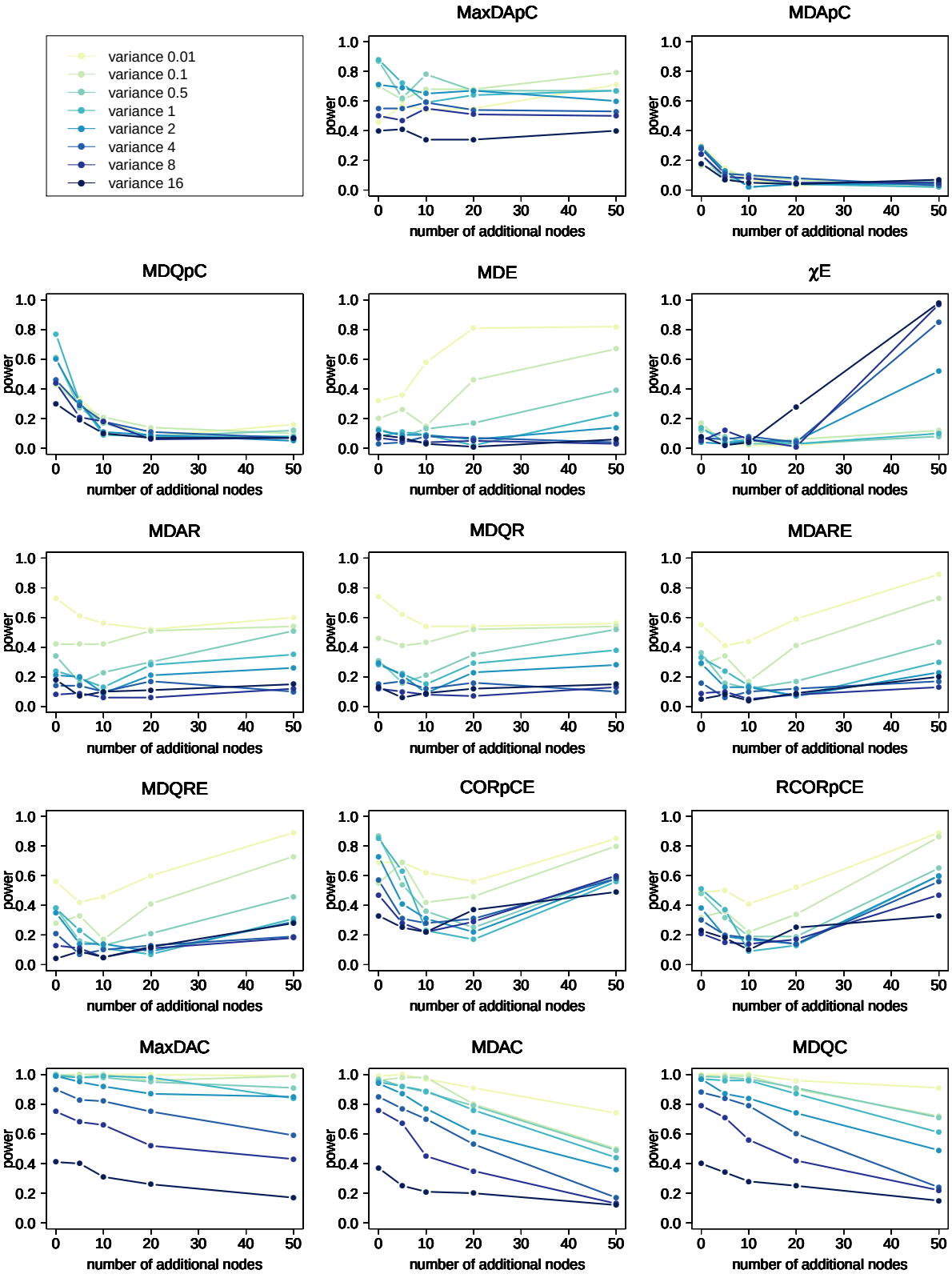

Figure 88: Proportion of rejected hypothesis for simulated setting of 180 and 20 samples and knockout of nodes PIP2, PKC, and PKA.

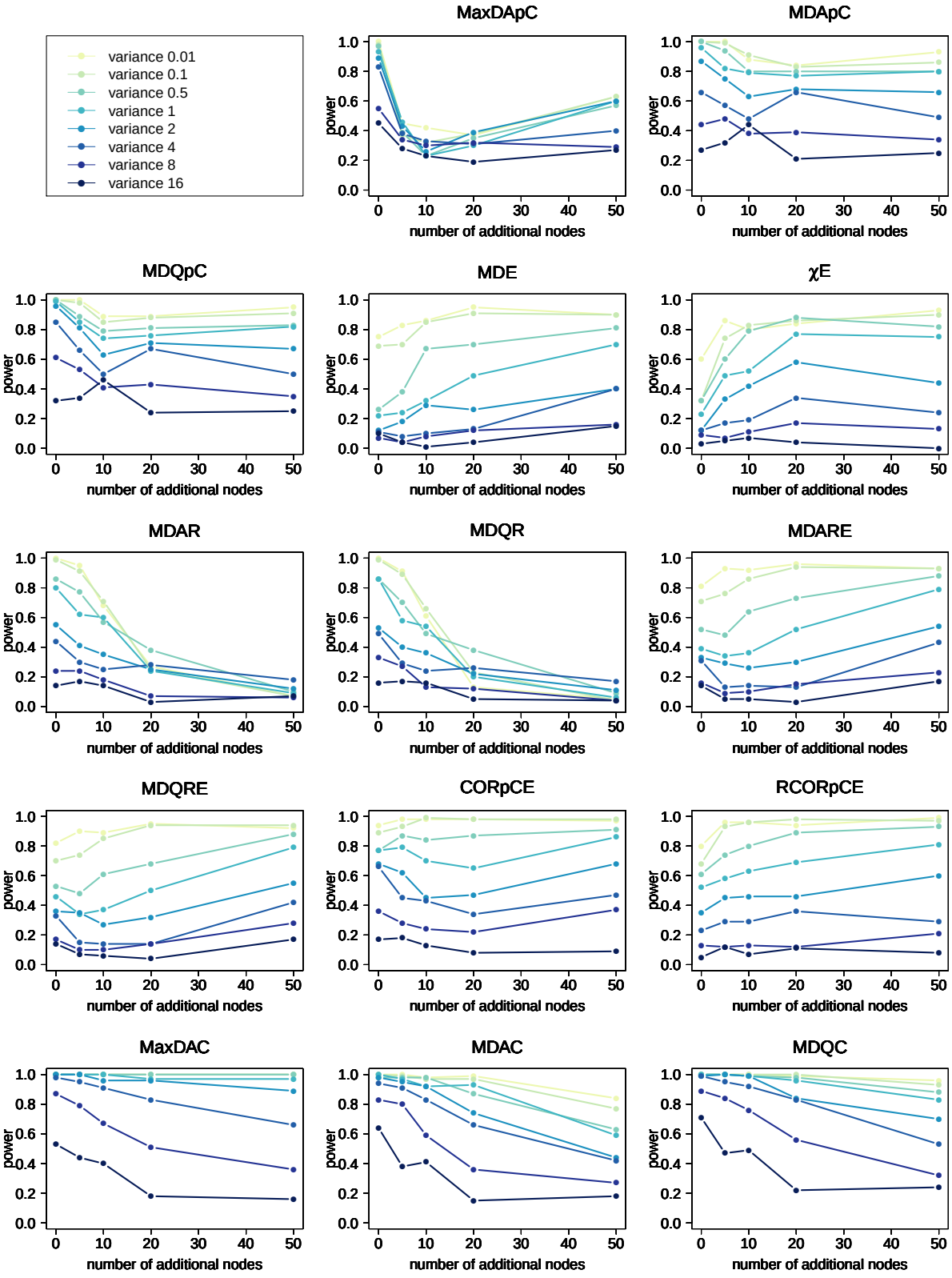

Figure 89: Proportion of rejected hypothesis for simulated setting of 50 samples per group and knockout of nodes PIP2, PKC, and PKA.

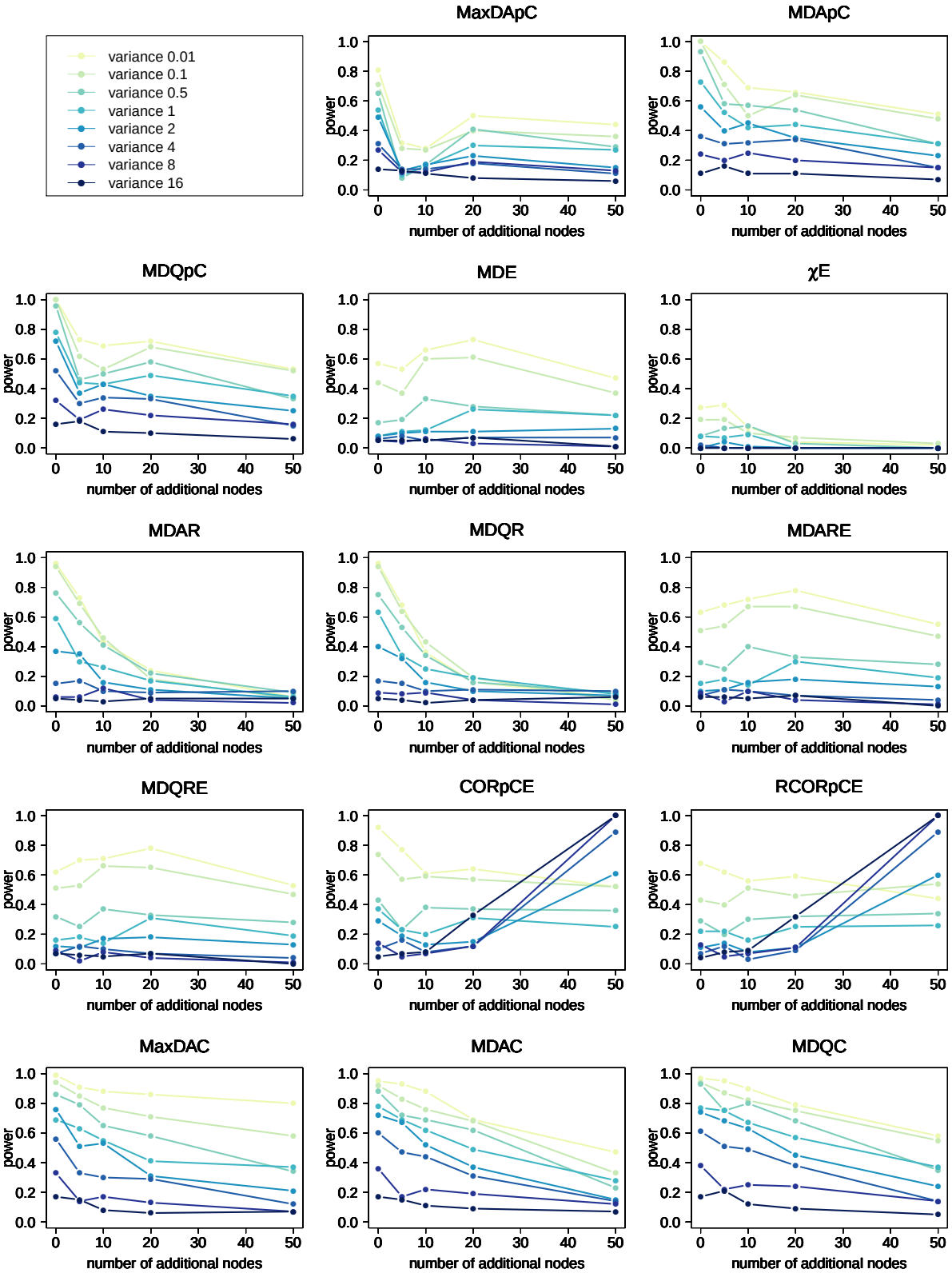

Figure 90: Proportion of rejected hypothesis for simulated setting of 20 samples per group and knockout of nodes PIP2, PKC, and PKA.

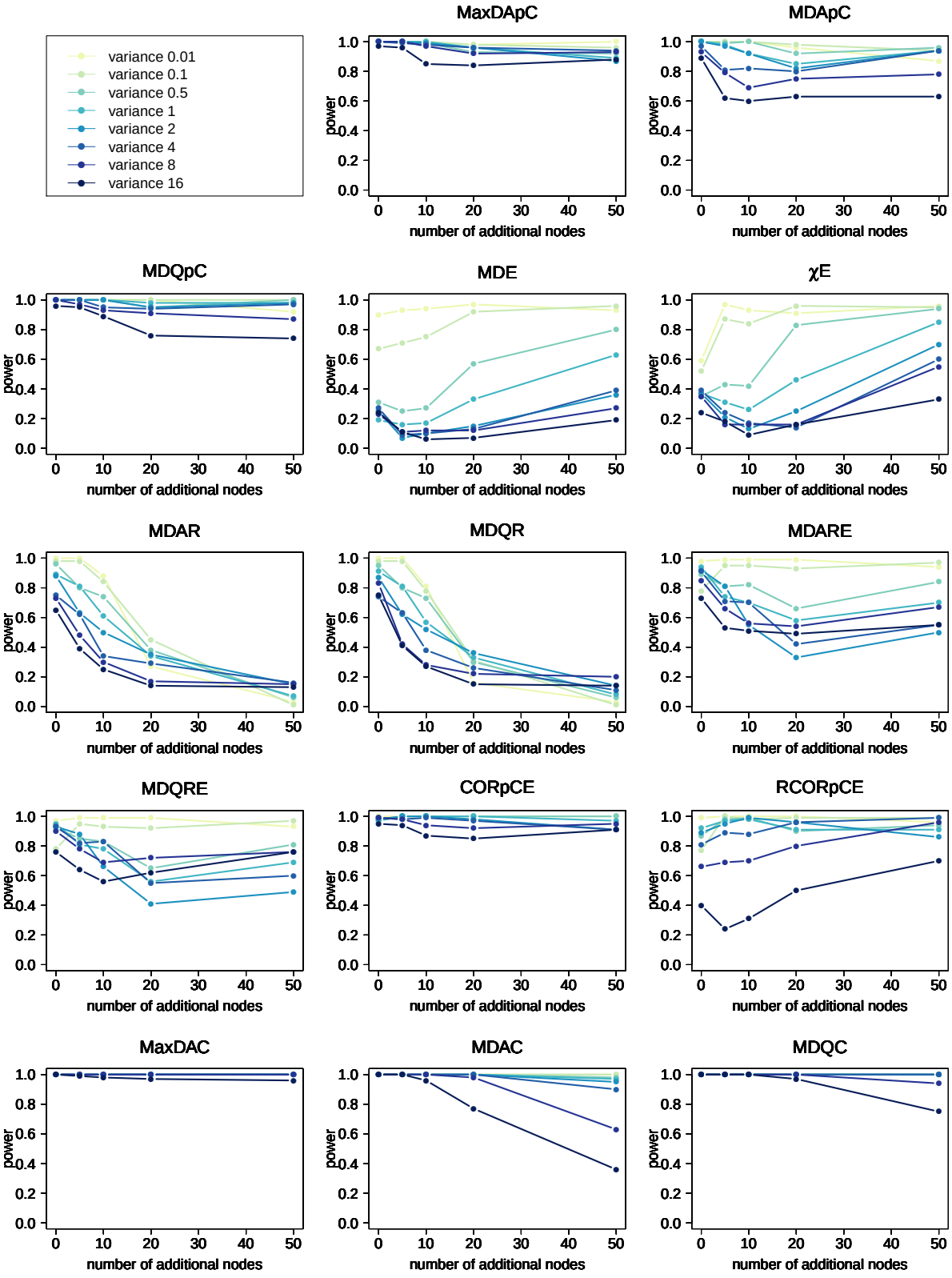

Figure 91: Proportion of rejected hypothesis for simulated setting of 200 samples per group and knockout of nodes PIP2, PKC, and PKA.

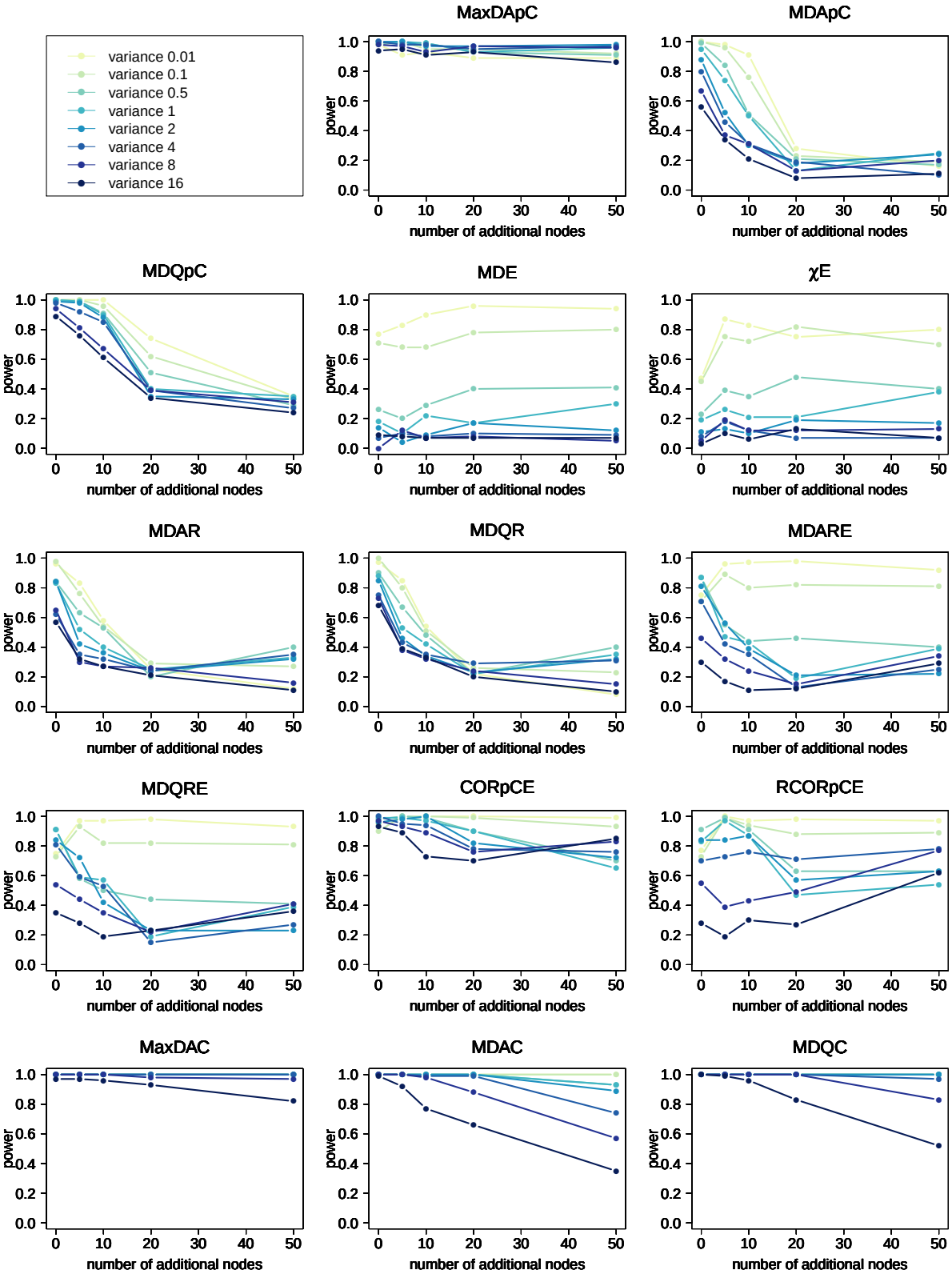

Figure 92: Proportion of rejected hypothesis for simulated setting of 300 and 100 samples and knockout of nodes PIP2, PKC, and PKA.

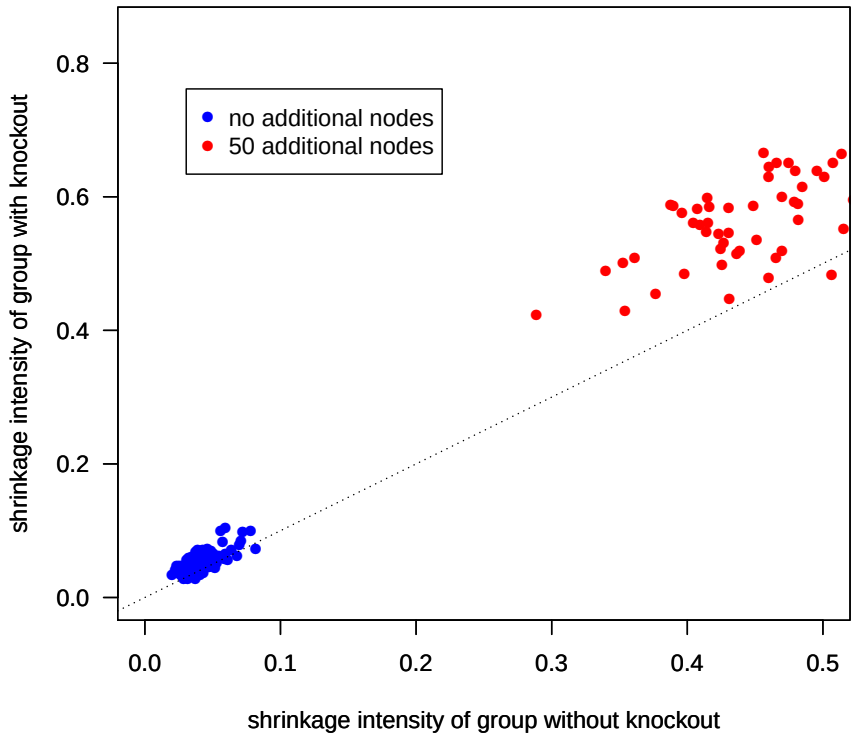

Figure 93: Comparison of shrinkage intensities in the scenario with 100 samples per group without (blue) and 50 additional nodes (red), where variance of the noise term is 1, and node PKC is knocked out.

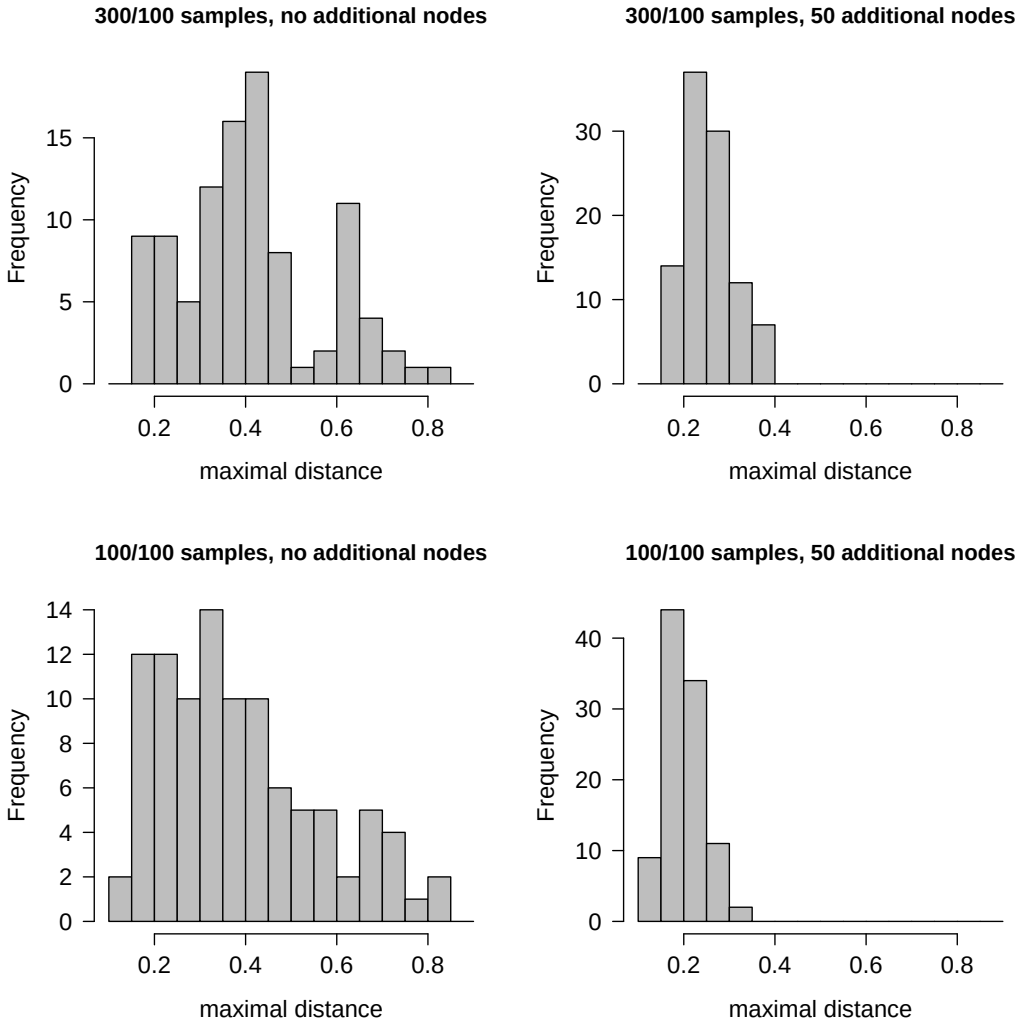

Figure 94: Histogramms of the maximal distances of partial correlations in the scenario with 300 samples in the first and 100 samples in the second group (top) and 100 samples in both groups (bottom) without (left) and 50 additional nodes (right), where variance of the noise term is 1, and node PKC is knocked out.

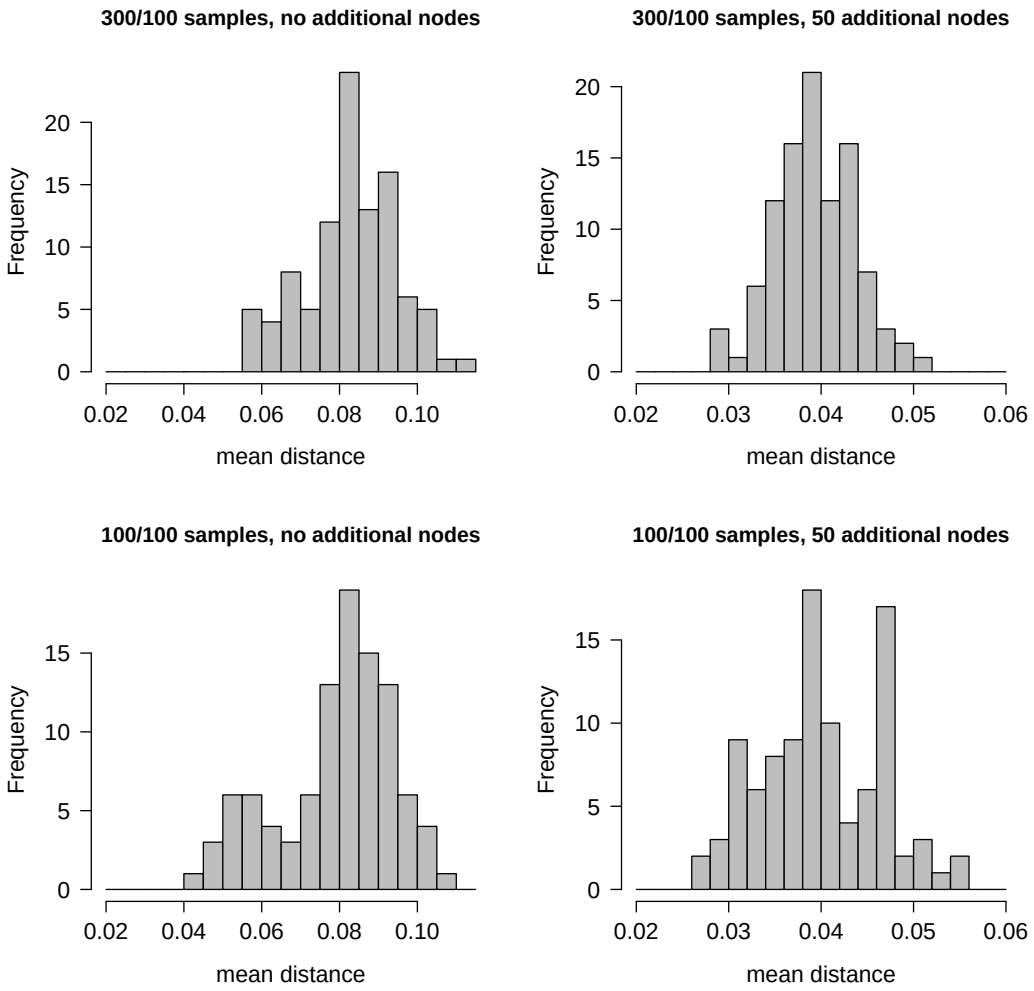

Figure 95: Histogramms of the mean distances of partial correlations in the scenario with 300 samples in the first and 100 samples in the second group (top) and 100 samples in both groups (bottom) without (left) and 50 additional nodes (right), where variance of the noise term is 1, and node PKC is knocked out.