Biomarkers, 2010; 15(8): 677–683

informa
healthcare

ORIGINAL ARTICLE

The impact of storage effects in biobanks on biomarker discovery in systems biology studies

Raji Balasubramanian[1], Laurin Müller[2], Karl Kugler[2], Werner Hackl[2], Lisa Pleyer[3], Matthias Dehmer[2], and Armin Graber[2]

[1]Division of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA, USA, [2]Institute for Bioinformatics and Translational Research, UMIT, Hall in Tirol, Austria and [3]Laboratory for Immunological and Molecular Cancer Research and Third Medical Department with Hematology, Medical Oncology, Hemostaseology, Rheumatology and Infectiology, Paracelsus Medical University Salzburg, Salzburg, Austria

Abstract
Sample handling and storage conditions in specimens frozen over long periods of time can severely impact marker levels. If laboratory technologies, practices and related protocols change over time, biomarker studies are potentially biased and report erroneous results. These issues and pitfalls are often overlooked in system biology studies using previously collected and stored materials, and are likely to be one notable cause for biomarker candidates failing to be validated. We present results from simulation studies quantifying the loss in statistical power to detect true biomarkers, due to diminishing concentration of analytes in samples subject to poor handling and storage conditions.

Keywords: *Computational biology; breast cancer; prostate cancer*

Introduction

Numerous putative biomarkers or panels of biomarker candidates for prognosis and diagnosis of cancer are reported in the literature. Nevertheless, only a very small number of these candidates are validated in subsequent studies (Chatterjee&Zetter 2005, Liotta&Petricoin 2008, Whiteley 2008). This high attrition rate can be attributed to inadequate planning of experiments using sound statistical principles of experimental design and the lack of high standardization and comprehensive quality control principles, thereby challenging the reproducibility of experimental findings in discovery studies (Hu et al. 2005, Pepe et al. 2001).

The use of banked specimens is attractive for biomarker discovery. Biological samples, including serum/plasma, DNA/RNA or tissues, have been collected, processed and stored in biobanks for many years in laboratories, clinical departments and hospitals. However, as sample handling and storage conditions in specimens frozen over long periods of time are neither consistent nor thoroughly evaluated and documented, many putative biomarker discoveries fail to be validated. Good laboratory practices and stringent quality control systems for sample handling and storage conditions in specimens frozen over long periods of time are crucial for biomarker discovery to yield results that are reproducible.

These concerns have resulted in several leading laboratories in human cancer biomarker research to mandate both experimental and preanalytical standards that guarantee uniformity in specimen collection, handling and storage (Banks et al. 2005, Raiet al. 2005, Schrohlet al. 2008). The UK and the Norwegian biobanks' sample handling and storage protocols for sample processing and archiving are examples of nationwide efforts to standardize some of these critical issues (Downey &Peakman 2008, Elliott &Peakman 2008, Rønningen et al. 2006). Ideally, biomarker discovery projects should be restricted to specimens collected by rigorous adherence to banking protocols. Additionally, diligent tracking and documentation of all preanalytical variables are essential, including full annotation of patient data and monitoring every single protocol step

Address for Correspondence: Armin Graber, Institute for Bioinformatics and Translational Research, UMIT, Eduard WallnoeferZentrum 1, A-6060, Hall in Tirol, Austria. Tel: +0043 50 8648 3803. Fax: +043 50 8648 67 3803. E-mail: armin.graber@umit.at

(Received 15 April 2010; revised 01 July 2010; accepted 25 July 2010)

ISSN 1354-750X print/ISSN 1366-5804 online © 2010 Informa UK, Ltd.
DOI: 10.3109/1354750X.2010.511265

from type and form of blood collection tube chosen, to storage temperature as changes of laboratory practices, technologies and related protocols over time might influence biomarker study outcome (Diamandis 2004a, b, Drake et al. 2004, Hill et al. 1992, Smetset al. 2004).

This work was motivated by our study of the effects of historically made changes in collection, handling and storage of samples on the resulting success of biomarker discovery studies. Change in the overall concentration levels of markers can be caused by various factors, such as the number of freeze–thaw cycles or chemical activity, which might modify the stored material. For instance, formalin fixation and paraffin embedding (FFPE) is the most commonly used method worldwide for tissue storage. FFPE preserves the tissue integrity but causes extensive damage to nucleic acids causing chemical changes and degradation in tissue DNA, RNA and protein (Farragheret al. 2008, von Ahlfenet al. 2007). In this analysis, we assume that there exists a gradual degradation of the concentration of individual biochemical analytes over time in biological specimens that are subject to poor storage and handling conditions. Time is not the only factor likely to influence changes in measured marker levels; nevertheless, it is a very important one. Other factors such as specimen size, or even conditions such as temperature and treatment of the samples prior to storage can influence the concentration as well. This degradation in samples specific to true biomarkers of the outcome of interest often translates to loss in statistical power and the failure to detect clinically relevant analytes. To quantify the loss in statistical power due to sample degradation over time, we present results from simulation studies of single variable and multivariate models commonly used for the analysis of data generated from high-throughput technologies in system biology studies. In all the analyses reported here, we assume that there are two groups being compared, referred to as 'cases' and 'controls'.

Methods

We first describe simulations assessing the effects of degradation for true univariate (or individual) biomarkers that are significantly associated with outcome. Subsequently, the consequences of degradation for multianalyte biomarker sets, i.e.panels of biomarkers that are predictive of outcome, are explored.

Notation

Let the random variable X denote the natural log-transformed levels of an analyte of interest. Then if X is a univariate biomarker of treatment, it satisfies:

$$\mu_x^{\text{Case}} \neq \mu_x^{\text{Control}}$$

where μ_x^{Case} denotes the expected value (mean) of X in the population of cases and μ_x^{Control} denotes the expected value (mean) of X in the population of controls. In our simulations, X was simulated according to a Gaussian distribution with variance 0.2 in each group. The difference between the expected value of X between the cases and controls (i.e. $|\mu_x^{Case} - \mu_x^{Control}|$) was varied between ln(1.5) and ln(1.0), resulting in a mean fold changes (MFC) (i.e. $|\mu_x^{Case}/\mu_x^{Control}|$) ranging between 1.5 and 1.

Univariate biomarkers of outcome study design and model

The goal of the simulation studies was to quantify the loss in statistical power for the detection of individual biomarkers, resulting from the degradation for biomarker levels in samples due to poor storage conditions.

Simulation study

One hundred datasets each consisting of 1000 analytes weregenerated for sample sizes of (1) 100 cases versus 100 controls and (2) 300 cases versus 300 controls. The percentage of true biomarkers in the dataset was assumed to be 1%, 2% or 3%, respectively. The statistical significance of each analyte was determined using a t-test. For each analyte, adjustment for multiple comparisons was carried out according to the methods for controlling the false discovery rate (FDR) (described in Benjamini &Hochberg 1995). For each of the conditions evaluated (sample size, MFC and proportion of biomarkers), the overall power was estimated as the proportion of all true biomarkers found to be statistically significant, when allowing an FDR of at most 5%.

Multianalyte biomarker set of outcome study design and model

The goal of the multivariate simulation study was to assess loss in statistical power to detect a multianalyte panel of biomarkers, resulting from the degradation for biomarker levels in samples due to poor storage conditions.

The simulations were conducted under the assumption that all analytes are independent, distributed according to a log-normal distribution characterized by a within-group biological variability of 0.2. We further assumed that profiling of 1000 analytes was conducted and either 1%, 2% or 3% were true biomarkers of the outcome of interest.Each analyte was assumed to have the identical value of MFC reflecting the ratio of the

average intensity of an analyte signal in the case group and the average intensity of the same analyte signal in the control group.

Each value of MFC was directly translated to an average case/control status classification accuracy or corresponding area under the curve (AUC) statistic associated with the multi-panel of a biomarker set that included either 10, 20 or 30 analytes at the particular value of MFC and sample size. The results shown are averages of 100 simulated datasets. For each simulated dataset, the class prediction algorithm PAM (Tibshiraniet al. 2002) was used to obtain the optimal biomarker set for predicting the case/control status of an individual subject. A *p*-value denoting the statistical significance of each classifier was calculated by comparing the prediction accuracy of the classifier with the distribution of prediction accuracy obtained under the null hypothesis (i.e. when all 1000 simulated analytes have a mean fold change of 1.0). Statistical power was then calculated as the proportion of *p*-values (among 100 simulated datasets) that were less than 0.05.

Results

The results described below quantify the loss in statistical power to detect individual biomarkers and multi-panel biomarker sets as a result of degradation of individual analyte levels in samples subject to poor sample handling and storage conditions. We assume that degradation of analyte levels in biological specimens result in diminishing values of MFC, reflecting smaller differences between mean biomarker levels in cases when compared with controls among samples subject to poor handling/storage. The results of the simulations of the single variable and multivariate models exploring storage effects are described below.

Simulations of storage effects on univariate biomarkers of outcome

To explore the impact on statistical power to discover true biomarkers, we conducted various simulation studies assuming sample sizes of 100 (300) cases and 100 (300) controls.We varied the MFC of true biomarkers between 1.5 and 1.0, to reflect diminishing differences between the mean analyte concentrations in cases compared with controls. The decreasing MFC levels reflect decreasing biomarker levels in frozen specimens over time. For example, a greater than 90% power to detect individual biomarkers with true population MFC of 1.35 or higher, translates to an effective power of only 50% assuming a MFC of 1.24 in the samples subject to degradation (see Figure 1 for sample size 100 vs 100). Similarly, a greater than 90% power to detect individual biomarkers with population MFC of 1.18 or higher, translates to an effective power of 50% assuming an MFC of 1.12 in samples due to degradation (see Figure 2 for sample size 300 vs 300).

Simulations of multianalyte biomarker set of outcome

We varied the MFC of each biomarker between 1.5 and 1.0, reflecting the effect of a decrease in biomarker levels in frozen specimens over time. Figures 3 and 4 present the relationship of the MFC of each individual component of a multianalyte biomarker set to the multivariate classifier's (1) AUC and (2) classification accuracy – here, the multianalyte biomarker set is assumed to comprise of 1%, 2% or 3% biomarkers, each with the same value of MFC, respectively. For instance, a classifier comprising 30 (10) biomarkers, where each individual biomarker has a MFC of 1.2 would result in an AUC of 0.82 (0.69), respectively (Figure 3). Likewise, a classifier comprising of 30 (10) biomarkers, where each individual biomarker has a MFC of 1.2 would result in an average classification accuracy of 0.85 (0.73), respectively (Figure 4).

The simulation results show that a greater than 99% power to detect a multianalyte biomarker set comprising 30 individual biomarkers, where each individual biomarker has a population MFC of 1.15 or higher, translates to a power of only 50% when the individual biomarker MFC is reduced to 1.11, due to degradation (see Figure 5 for sample size 100 vs 100). Similarly, a greater than 99% power to detect a panel of 30 biomarkers, where each individual biomarker has a population MFC of 1.08 or higher, translates to a power of only 50% assuming

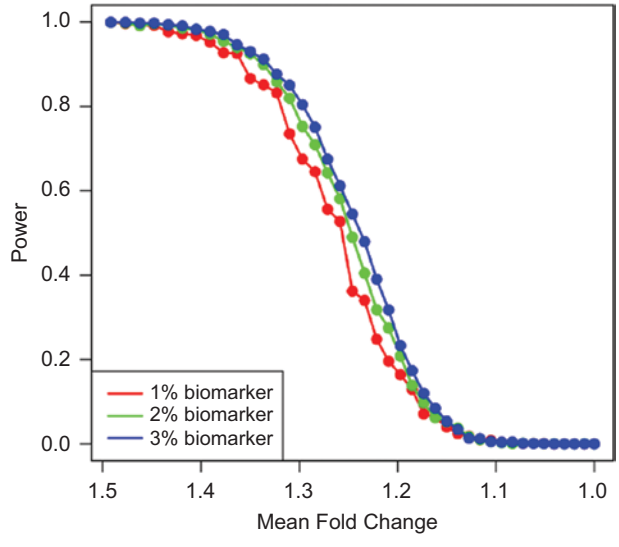

Figure 1. Mean fold change (i.e. ratio of mean intensity in cases to that in controls) of individual biomarker versus proportion of biomarkers detected as statistically significant based on a sample size of 100 cases and 100 controls.

a MFC of 1.06 in the sample due to degradation (see Figure 6 for sample size 300 vs 300).

Discussion

All previously reported results are based on statistical models that the biomarkers are assessed in case–control data. In this respect, estimates of the variance of the measured markers and expected effect sizes are central assumptions. In general, important components of variation are the between-subject, the within-subject and the lab/measurement variation. If these are large, then it frequently becomes very difficult to determine that reasonable differences between cases and controls are statistically significant. The distribution of coefficients of variation (CV), a normalized measure of dispersion, of technical replicates depends typically on the analytical platform (e.g. lipid or polar

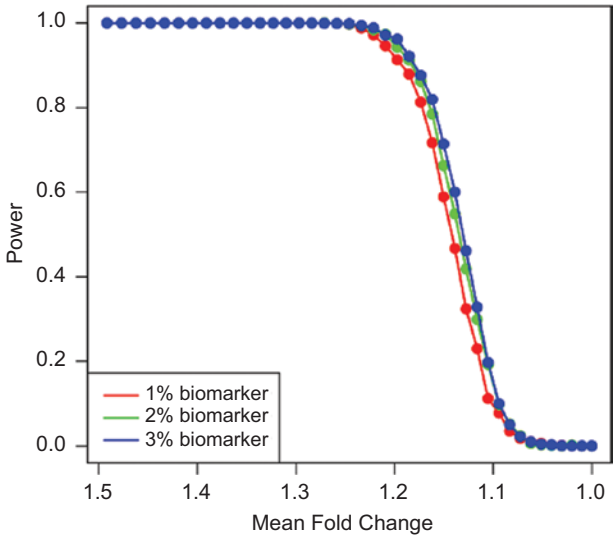

Figure 2. Mean fold change (i.e. ratio of mean intensity in cases to that in controls) of individual biomarker versus proportion of biomarkers detected as statistically significant based on a sample size of 300 cases and 300 controls.

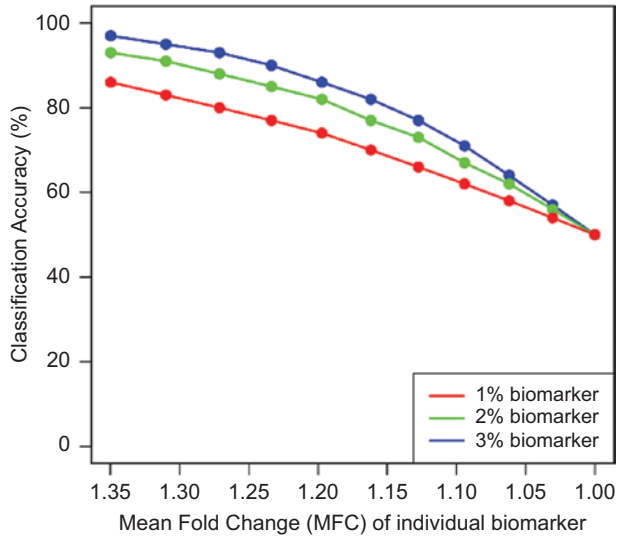

Figure 4. The relationship of the mean fold change of individual biomarker (in multianalyte set) with the classification accuracy of a classifier based on the multianalyte biomarker set.

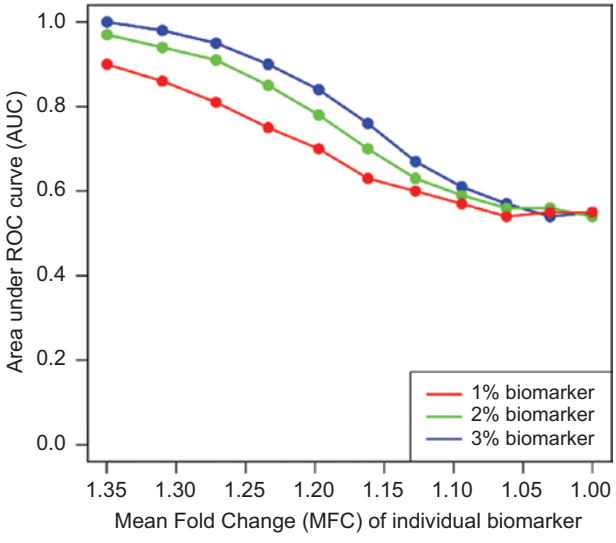

Figure 3. The relationship of the mean fold changeof individual biomarker (in multianalyte set) with the area under the curve of a receiver-operating characteristic (ROC) curve based on the multianalyte biomarker set.

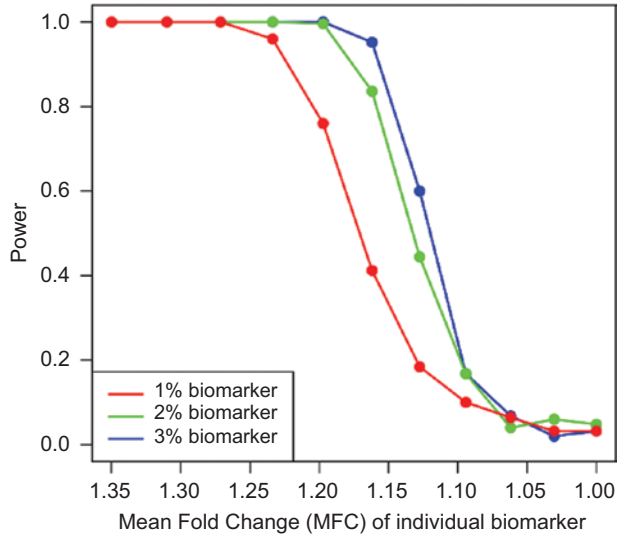

Figure 5. Power to detect a multianalyte biomarker set as a function of mean fold change of each individual biomarker that comprises the multianalyte set.Sample size is assumed to be 100 cases and 100 controls.

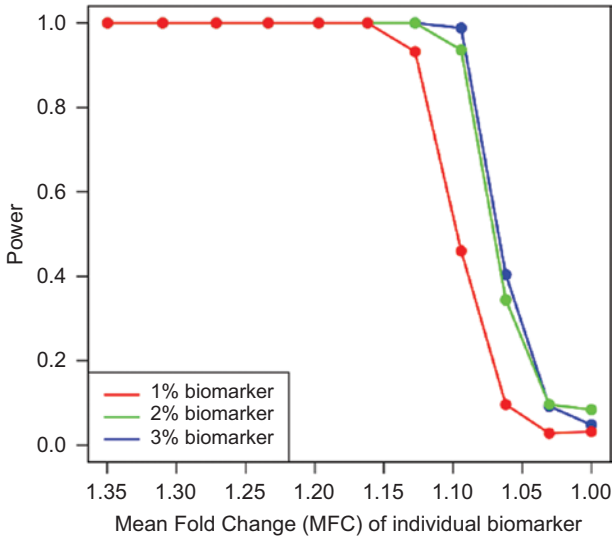

Figure 6. Power to detect a multianalyte biomarker set as a function of mean fold change of each individual biomarker that comprises the multianalyte set.Sample size is assumed to be 300 cases and 300 controls.

liquid chromatography (LC)-mass spectrometry (MS), gas chromatography-MS, LC-MS/MS), profiling strategies (e.g. non-targeted vs targeted) and technical approaches such as scan mode (e.g. product ion, neutral loss or MRM); however, it is also influenced by other factors, e.g. sample material type such as plasma or tissue. The majority of analytes based on technical replicates should report CVs of 10% or lower on a well-performing research platform.

The simulations were conducted under the assumptions that all analytes are distributed according to a log-normal distribution and are characterized by a within-group biological variability of 0.2, which was observed with a representative platform in previous system biology studies (McBurneyet al. 2009, van der Greefet al. 2007). Furthermore, our model assumes that alteration of individual marker levels caused, for instance, by degradation processes leads to a decrease of marker levels over time, which is reflected by lower MFC values. We recently reported the impact of storage time on biomarker levels in frozen serum and measured the concentration of the two biomarkers, CA 15-3 and CA125, in samples that were collected between April 1995 and April 2001 and stored at -70ºC (Kugleret al. 2010). This simple experimental estimation shows that the predominance of biomarker levels using these tests increased over time, which might be related to the used blood collection tubes and caps, and potential leakage. Several groups have reported the confounding influence that specimen collection devices may have on proteomic measurements (Diamandis 2004a, b, Hill et al. 1992, Lippi et al.2005, 2006, Pilny et al. 2006). Nevertheless, we decided to demonstrate the effects of a mere decline of marker levels in frozen specimens over time as this scenario seems to reflect the more likely storage effect. However, all presented graphs can also be alternatively interpreted for a marker level increase or increase of markers above a certain MFC by reversing the *x*-axis.

In conclusion, we have demonstrated the drastic reduction in statistical power resulting from decreasing concentrations of individual biomarkers over long periods of specimen storage. In our analyses we assumed that poor handling and storage conditions result in uniform decline in biomarker levels over time in both cases and controls, resulting in smaller differences in mean biomarker levels between the cases and controls in compromised sample sets. In contrast, changes in sample handling procedures over time can also result in preferential degradation of analytes in some sample sets over time. In these situations, spurious biomarkers can be detected, reflecting sample handling effects rather than true biological effects.

Our model, although simplistic, allowed us to investigate the effects of marker level change on the outcome of biomedical studies in a generic manner. We did not explicitly identify and model specific sources of degradation and did not limit the interpretation to specific decay functions for such a change, which allows a more generic study of collective changes due to time, freeze-and-thaw circles, chemical activities or any other degradation source. Furthermore, we assumed that the distributions of the 1000 features were independent. However, this assumption might not be valid in applications due to the inherent dependence of biomarkers that belong to similarly acting biological pathways. To model this dependence between biomarkers, the previously described data generation model was generalized to incorporate a correlation of 0.5 between the ten biomarkers in the cases. The 990 noise features were assumed to be independent in both cases and controls as were the ten biomarkers in the control group. As expected, the statistical power was lower among all classifiers when compared with the setting in which all biomarkers were assumed to be independent among both cases and controls (Guoet al. 2010). In this study, for computational simplicity, we decided not to include the inherent dependence of biomarkers in the model. A more complex model would not have changed any of the major conclusions of this article.

Additionally, we used MFC as a calculated measure in the simulations. MFC changes are frequently used in omic experiments, where many assays only allow relative quantitation (McBurney et al. 2009, van der Greefet al. 2007). Another simplistic assumption for our simulation was that we modelled all the markers to have a variability of 0.2. Ananalysis of the effect of higher or lower variability remains to be investigated in future studies.

The development and validation of a new clinical assay or biomarker ensures that a test is reproducible, consistently meets defined performance characteristics and specified clinical utility. However, this information can only be efficiently obtained through the stringent and consistent use of experimental design, standardization and quality control principles. The organization and documentation of this work through a quality system is both mandated and practical. Obviously, the quality of data obtained is significantly influenced by the quality of the specimen used to generate the data. In this respect, an important source of variability and bias relates to changes in sample collection, handling and storage conditions over time and sample age, which may result from serum degradation related to storage time or the number of freeze–thaw cycles.

Without good laboratory and banking practices, the scientific community will continue to waste time and money in systems biology studies, and report distorted results that are misleading and cannot be reproduced in blinded studies. Moreover, scientists have an obligation to scrutinize and cross-examine experimental designs and quality control principles in order to minimize the exploitation of any human effort and resources, even when ethical standards are considered in research plans.

Acknowledgements

This work was supported by the COMET Center ONCOTYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the TirolerZukunftsstiftung (TZS) and the State of Styria represented by the Styrian Business Promotion Agency (SFG) (and supported by the University for Health Sciences, Medical Informatics and Technology and BIOMAX Informatics AG).

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

Banks RE, Stanley AJ, Cairns DA, Barrett JH, Clarke P, Thompson D, Selby PJ. (2005). Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clin Chem*51:1637–49.

Benjamini Y, Hochberg Y. (1995).Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat SocSer B (Methodological)* 57:289–300.

Chatterjee SK, Zetter BR. (2005). Cancer biomarkers: knowing the present and predicting the future. *Future Oncol*1:37–50.

Diamandis EP. (2004a). Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J Natl Cancer Inst*96:353–6.

Diamandis EP. (2004b). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics*3:367–78.

Downey P, Peakman TC.(2008). Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles.*Int J Epidemiol 37* (Suppl. 1): i46–50.

Drake SK, Bowen RA, Remaley AT, Hortin GL. (2004).Potential interferences from blood collection tubes in mass spectrometric analyses of serum polypeptides.*Clin Chem*50:2398–401.

Elliott P, Peakman TC. (2008). The UK Biobank sample handling and storage protocol for the collectionprocessing and archiving of human blood and urine. *Int J Epidemiol*37:234–44.

Farragher SM, Tanney A, Kennedy RD, Harkin PD. (2008). RNA expression analysis from formalin fixed paraffin embedded tissues. *Histochem Cell Biol*130:435–45.

Guo Y, Graber A, Mc Burney RN, Balasubramanian R. (2010). Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *BMC Bioinformatics* under revision.

Hill BM, Laessig RH, Koch DD, Hassemer DJ. (1992). Comparison of plastic vs. glass evacuated serum-separator (SST) blood-drawing tubes for common clinical chemistry determinations. *Clin Chem*38:1474–8.

Hu J, Coombes KR, Morris JS, Baggerly KA. (2005). The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic*3:322–31.

Kugler K.,Hackl W., Müller L., Graber A. & Pfeiffer R. (2010).Quantification of Storage Effects in Biobanks and the inflicted Bias for Biomarker Discovery Studies. *in preparation*.

Liotta LA, Petricoin EF. (2008). Putting the 'Bio' back into biomarkers: orienting proteomic discovery toward biology and away from the measurement platform. *Clin Chem*54:3–5.

Lippi G, Montagnana M, Salvagno GL, Guidi GC. (2006).Interference of blood cell lysis on routine coagulation testing. *Arch Pathol Lab Med*130:181–4.

Lippi G, Salvagno GL, Montagnana M, Brocco G, Guidi GC. (2005). Influence of short-term venous stasis on clinical chemistry testing. *Clin Chem Lab Med*43:869–75.

Mcburney RN, Hines WM, Von Tungeln LS, Schnackenberg LK, Beger RD, Moland CL, Han T, Fuscoe JC, Chang CW, Chen JJ, Su Z, Fan XH, Tong W, Booth SA, Balasubramanian R, Courchesne PL, Campbell JM, Graber A, Guo Y, Juhasz PJ, Li TY, Lynch MD, Morel NM, Plasterer TN, Takach EJ, Zeng C, Beland FA. (2009). The liver toxicity biomarker study: phase I design and preliminary results. *ToxicolPathol*37:52–64.

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. (2001).Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 93: 1054–61.

Pilny R, Bouchal P, Borilova S, Ceskova P, Zaloudik J, Vyzula R, Vojtesek B, Valik D. (2006).Surface-enhanced laser desorption ionization/time-of-flight mass spectrometry reveals significant artifacts in serum obtained from clot activator-containing collection devices. *Clin Chem* 52:2115–16.

Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, Mehigh RJ, Cockrill SL, Scott GB, Tammen H, Schulz-Knappe P, Speicher DW, Vitzthum F, Haab BB, Siest G, Chan DW. (2005). HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics* 5:3262–77.

Rønningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KK, Hovengen R, Haugen M, Nystad W, Magnus P, Hoppin JA. (2006). The biobank of the Norwegian Mother and Child Cohort Study: a resource for the next 100 years. *Eur J Epidemiol* 21: 619–25.

Schrohl AS, Wurtz S, Kohn E, Banks RE, Nielsen HJ, Sweep FC, Brunner N. (2008).Banking of biological fluids for studies of disease-associated protein biomarkers.*Mol Cell Proteomics* 7:2061–6.

Smets EM, Dijkstra-Lagemaat JE, Blankenstein MA. (2004). Influence of blood collection in plastic vs. glass evacuated serum-separator tubes on hormone and tumour marker levels. *Clin Chem Lab Med* 42:435–9.

Tibshirani R, Hastie T, Narasimhan B, Chu G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *ProcNatlAcadSci U S A* 99:6567–72.

Van Der Greef J, Martin S, Juhasz P, Adourian A, Plasterer T, Verheij ER, Mcburney RN. (2007). The art and practice of systems biology in medicine: mapping patterns of relationships. *J Proteome Res*6:1540–59.

Von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. (2007).Determinants of RNA quality from FFPE samples.*PLoS One* 2:e1261.

Whiteley G. (2008).Bringing diagnostic technologies to the clinical laboratory: rigorregulationand reality.*Proteomics Clin Appl*2:1378–85.