EMBO reports

meeting report

From microarray data to results

Workshop on Genomic Approaches to Microarray Data Analysis

Thomas Schlitt[1,2+] & *Patrick Kemmeren*[1,3]

[1]EMBL-EBI, Cambridge, UK, [2]British Antarctic Survey, Cambridge, UK, and [3]Division of Biomedical Genetics, UMC Utrecht, Utrecht, The Netherlands

This European Science Foundation (ESF) exploratory workshop was held at the Centro Nacional de Investigaciones Oncologicas (CNIO) in Madrid on 30 and 31 October 2003, and was organized by J. Dopazo, R. Díaz-Uriarte and A. Brazma.

Keywords: data analysis; genomics; microarray; statistics

EMBO reports (2004) **5**, 459–463. doi:10.1038/sj.embor.7400156

Introduction

The main focus of this workshop was to discuss the challenges involved in the collection, storage and analysis of the large amounts of biological data that are being produced by microarray technology.

[1]European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]British Antarctic Survey, Natural Environment Research Council, High Cross, Madingley Road, Cambridge CB3 0ET, UK
[3]Genomics Laboratory, Division of Biomedical Genetics, UMC Utrecht, PO Box 85060, 3508 AB Utrecht, The Netherlands
[+]Corresponding author. Tel: +44 1223 221656; Fax: +44 1223 221259; E-mail: schlitt@ebi.ac.uk

Submitted 16 February 2004; accepted 26 March 2004; published online 23 April 2004

Microarray experiments can be conceptually subdivided into material- and data-processing steps. During material processing, important information needs to be recorded, such as array design, experimental conditions and sample treatment, to enable meaningful data analysis and biological interpretation. This workshop concentrated on the subsequent microarray data processing, which can be further divided into data preprocessing, such as normalization and filtering, data-analysis steps and the biological interpretation of the results.

The first steps in microarray data preprocessing involve image scanning, and include spot finding and the selection of good quality spots. Next, data-normalization steps are necessary to correct unavoidable experimental variations, such as differences in sample preparation, dye incorporation and hybridization efficiencies. These variations are not owing to differences in gene expression in the original samples and, therefore, need to be corrected before data analysis can be carried out. Such analysis might include various methods to identify genes that are differentially expressed or conditions (cell-culture treatments, diseases and so on) that result in similar changes in gene expression. Some normalization or data-analysis methods require special arrangements, such as a particular array design. Therefore, both material- and data-processing steps need to be considered at the early stages of a microarray experiment. The biological interpretation of the data is facilitated by various tools, which place the analysis results into context with existing biological knowledge, such as the scientific literature or sequence data. Efforts to unify and standardize the way in which information is recorded are making the interpretation of large-scale experiments easier. Finally, the integration of biological information from various sources, such as large-scale data sets produced by various experimental techniques, provides a valuable platform for the exploration of regulatory networks. All of these topics were discussed during the workshop and a summary of the research that was presented is given here.

Microarray data sharing

It is important that all of the information about a microarray experiment is recorded systematically, so that meaningful data sets can be generated. A. Brazma (Cambridge, UK) showed that microarray

data sets can be complex, so it is of particular importance to establish standards to enable microarray data to be shared efficiently. Such a standard has been defined by the Microarray Gene Expression Data Society (MGED; http://www.mged.org) and is now widely accepted by biological journals (Brazma *et al*, 2001). The European Bioinformatics Institute (EBI) offers a public data repository known as ArrayExpress that conforms to the MGED requirements and stores information about microarray experiments, including material-processing aspects such as experimental design, sample treatment and array designs. Furthermore, the EBI provides two web-based tools that allow scientists to analyse microarray data (Expression Profiler) and to submit microarray data to ArrayExpress (MiameExpress). Brazma also discussed the special arrangements that have been made for the submission of extremely large microarray data sets.

Experimental design and normalization

Normalization is a particular type of preprocessing that is applied to correct systematic variations both in and between data sets, such as differences in labelling efficiencies. Choosing the appropriate experimental and array design facilitates data normalization and further downstream analysis. P. Kemmeren (Utrecht, The Netherlands) presented a normalization method to accurately determine differential gene-expression levels using external controls. Most normalization methods assume that the messenger RNA (mRNA) expression levels of only a few genes change in each condition, or that changes in mRNA content are balanced (that is, a similar number of genes are upregulated and downregulated in each particular condition). Changes are calculated relative to the majority of transcripts but if global shifts in mRNA occur, such as in the yeast stationary phase, these methods can be misleading (Fig 1). Kemmeren showed that global changes in mRNA levels can be monitored more accurately with the use of external RNA controls, such as *Bacillus subtilis* mRNA, which are added to the samples in known concentrations (van de Peppel *et al*, 2003).

High background noise in the measurements can cause further problems. For low-intensity signals, background noise can be close to the signal intensity itself, leading to increased variance that confounds the detection of gene-expression changes for weakly expressed genes. A. von Heydebreck (Berlin, Germany) presented a computational method called 'vsn' to stabilize the variance across the intensity range. This variance-stabilization method uses the dependence between the variance and mean intensities to derive a transformation such that the variance becomes approximately independent of the mean intensities (Huber *et al*, 2002; an implementation of vsn is available online as an R package at http://www.dkfz.de/abt0840/whuber). The transformed ratio provides a more reliable measure for differential gene expression that can be used in downstream analyses regardless of the intensity range.

Data analysis, clustering and gene selection

Many different data-analysis methods can be applied to a microarray data set after the normalization step, depending on the particular questions being studied. The Gene Expression Pattern Analysis Suite (GEPAS), which was developed by J. Herrero (Madrid, Spain) and co-workers (Herrero *et al*, 2003), contains many tools to identify functionally related genes. It allows preprocessing of the data, execution of pairwise comparisons, gene

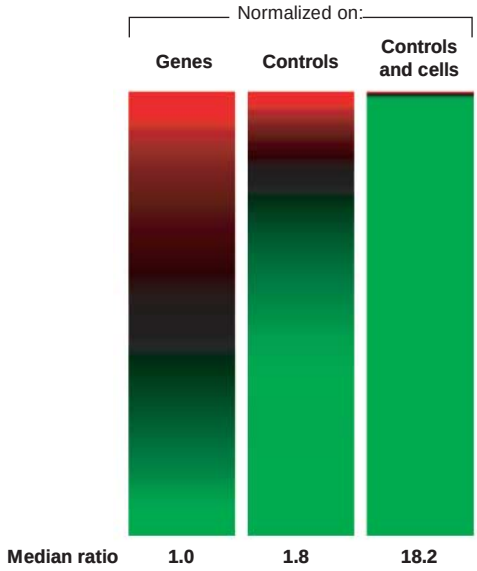

Fig 1 | The perception of changes in gene expression can depend on the normalization method used. The same data set and normalization method was used based on endogenous genes (left column), external controls (middle column), and external controls and cell count (right column). Upregulated genes are shown in red and downregulated genes in green. Numbers below the bars indicate the median change of all genes after normalization (as a ratio).

selection, unsupervised clustering and molecular classification, and is linked to the web-based tool FatiGO (http://fatigo.bioinfo.cnio.es; see later) for gene-annotation retrieval.

An important task in data analysis is the detection of differentially expressed genes. However, multiple tests for differentially expressed genes raise the probability of making false discoveries, as is the case for multiple tests in general, and this problem increases with the number of comparisons made. For example, an error rate of 5% might be acceptable in an individual test, but if 100 such tests are performed, there will probably be five false-positive results and, furthermore, their identity will be unknown. A solution could be to choose smaller significance levels for the individual tests to reduce the total number of false positives, but this is not feasible for microarray experiments given the large number of genes and limited precision of the technology. Different approaches can be taken to decide which tests provide acceptable error rates. S. Dudoit (Berkeley, USA) proposed general multiple-testing procedures for controlling false-positive rates, such as the generalized family-wise error rate (FWER). Using this procedure, the probability of finding at least one false positive is minimized. This new framework covers a broad range of testing problems that cannot be handled by traditional procedures, such as tests concerning parameters in survival models, pairwise correlations and Gene Ontology (GO) annotation. In a different approach, Y. Benjamini and A. Reiner (Tel Aviv, Israel) proposed the use of false-discovery rate (FDR) to control the number of false positives (Reiner *et al*, 2003). The FDR is defined as the expected proportion of false positives among the rejected hypotheses; that is, the proportion of falsely discovered differentially expressed genes. The expected proportion of false positives is estimated from running the same analysis on randomized data. The FDR is less conservative

than the FWER. The FWER is more appropriate for analyses in which a single false positive is unacceptable, such as comparing various drug treatments with a control. By contrast, the FDR is more applicable in screens for candidate genes, in which a certain proportion of false positives among the discovered genes is acceptable.

M. van der Laan (Berkeley, CA, USA) proposed a new method to find the optimal predictor for multivariate regression analysis, which involves predicting the outcome of a certain experiment on the basis of a number of variables, such as gene-expression levels. The deletion/substitution/addition algorithm (DSA) minimizes the residual sum of squared errors over a subset of basis functions. After training, the algorithm can be used as a black-box algorithm for multivariate regression; for example, to detect transcription-factor binding sites using yeast gene-expression data. G. Valentini (Milan, Italy) proposed a method for cancer classification using support vector machines (SVMs) as a classifier (Valentini, 2003). As not all genes are relevant for distinguishing normal from malignant tissues, feature-selection methods are used to select only those genes that are necessary for correct classification, therefore reducing the effect of background noise on the outcome of the prediction. Preliminary results indicate that the low-bias bagging (Lobag) approach used by Valentini and colleagues in association with feature selection outperforms other SVMs for the detection of normal and malignant tissues.

The work of R. Díaz-Uriarte (Madrid, Spain) addresses the identification of molecular signatures from biological data. Gene-expression signatures are sets of genes that are co-ordinately expressed and are related to the phenotypic condition. Most existing models for the identification of molecular signatures fail to address both of these requirements. An alternative approach is to identify a seed gene with good predictive abilities and then to iteratively look for groups of genes that are highly correlated both with the seed gene and among themselves, which also have good predictive abilities. Genes are eliminated from the group if they show only a small correlation with the seed gene or do not improve the prediction accuracy. According to tests with simulated and real data sets, Díaz-Uriarte showed that the performance of this algorithm is comparable with other state-of-the-art methods. The features learned (for example, the identification of a group of predictive genes) are interpretable, and the algorithm can be easily applied to other classifiers and other types of dependent variable (for example, survival analysis).

Information mining and automatic annotation methods

It is important for successful data mining to keep the information about genes that are represented on a microarray up to date, but this can be difficult for commercially produced arrays. J. De Las Rivas (Salamanca, Spain) introduced the tool Dynamic Annotation of GeneChip probe sets from Affymetrix (DAGA) for the identification and annotation of genes that are included on the oligonucleotide arrays from Affymetrix. The program generates a consensus sequence on the basis of the combination of the probes that form each set, and uses this consensus to search with BLAST for homologous sequences in mouse or human genome databases. The tool allows the validation of each probe set and their reassignment to genes, therefore keeping the annotations up to date with the information in the sequence databases.

Another problem faced by many scientists after performing cluster analysis of microarray data is how to identify biologically meaningful clusters and put them into context with the published literature to develop new hypotheses. This task is confounded by problems at several levels. Information retrieval—for example, locating all of the articles about one specific gene—can be difficult because many genes have several names. Unfortunately, in some cases, the same gene name refers to several different genes and, similarly, some acronyms are used to abbreviate several different terms. J. Tamames (Madrid, Spain) discussed the available text-mining tools, including TextDetective, which supports the retrieval of articles on particular genes, proteins, drugs or diseases, and TextMiner, which identifies the most relevant information from a set of articles. These tools are based on statistical methods, such as comparing word frequencies between articles of interest and all other articles. A. Valencia (Madrid, Spain) proposed integrating literature information into clustering analysis of microarray data to identify biologically meaningful clusters (Blaschke *et al*, 2001). He described one such approach, which integrates the clustering of expression profiles with the text-mining system Gene Expression Information System for Human Analysis (GEISHA). Valencia also pointed out that text-mining competitions are held to compare the performance of automated literature-mining systems (for further details, see http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04).

J. Dopazo (Madrid, Spain) described the web-based tool FatiGO, which supports the biological interpretation of clusters on the basis of the incorporation of biological knowledge derived from GO (Ashburner *et al*, 2000). GO is a hierarchical system of controlled vocabularies that is used by many biological databases to annotate proteins in a standardized hierarchical fashion. FatiGO finds GO terms that describe a group of genes with respect to a reference set, such as the remainder of the genome, and estimates the significance of the results (Al-Shahrour *et al*, 2004). J. Komorowski (Uppsala, Sweden) presented a method to use microarray data to infer the participation of genes in biological processes (Lagreid *et al*, 2003). Templates, such as constant expression, increasing expression or decreasing expression, are used to describe the expression patterns. Time-series expression profiles are divided into possible subintervals, each of which is assigned to an expression template using Boolean reasoning. On the basis of the expression templates, genes are assigned to a biological process using a 'guilt-by-association' approach. For a given profile, all applied rules are examined (over all possible subintervals) and a functional annotation is assigned on the basis of a majority vote.

In addition to looking for literature information on clusters, shared regulatory mechanisms can be explained by looking for *cis*-regulatory sequence motifs. This is performed under the assumption that coexpression at the transcriptional level is associated with transcriptional coregulation, which, in turn, is reflected at the sequence level by the presence of transcription-factor binding sites. Y. Moreau (Leuven, Belgium) presented a method for discovering *cis*-regulatory motifs using microarray data to identify known motifs using position-weight matrices and unknown motifs using a Gibbs motif sampling method, which searches for the statistically most probable motifs in a set of nucleotide sequences, and can find the optimal width and number of these motifs in each sequence (Lawrence *et al*, 1993). However, individual motifs often cannot explain the patterns in gene-expression data, and combinations of the motifs might be more appropriate; for example, in cases of synergistic effects between transcription factors. A genetic algorithm can therefore be used to search efficiently through all possible combinations of motifs. Although this approach works well for yeast, there are

problems with large non-compact genomes in which too many potential motifs are being found. The results for these genomes can be improved by restricting the analysis to evolutionarily conserved regions. All of these methods are part of the Java application TOUCAN, which is being developed by Y. Moreau and co-workers (Aerts *et al*, 2003).

Gene networks

The availability of high-throughput technologies makes it possible to explore large-scale regulatory networks, but also highlights the limitations of the existing modelling techniques and data sets. New approaches are necessary to analyse gene networks on a genome-wide scale. One particular problem that was addressed by several speakers is the integration of high-throughput data from heterogeneous sources, such as data on gene expression, mutant phenotypes and protein complexes. These integrated data sets form the basis for the subsequent analysis of gene networks. R. Shamir's group (Tel Aviv, Israel) is focusing on the modelling and analysis of networks that involve transcription regulation and metabolism. General steps in network inference include the definition of a class of possible networks and a scoring function, which scores how well a solution fits the data. Ideally, the search algorithm should find all possible solutions to the problem, but frequently the search space is too big. Shamir therefore proposed that a partially known network should be taken and the model should be refined by adding further levels of detail. He presented a model for lysine biosynthesis in the yeast *Saccharomyces cerevisiae* on the basis of data from the literature. Hypotheses about the regulation of lysine biosynthesis derived from this model were then compared with biological data to validate the model. He also presented a computational model in which changes in mRNA concentration are computed on the basis of transcription-factor concentrations, transcription-factor–DNA affinity and DNA signals in the promoter (Tanay & Shamir, 2003). Y. Barash (Jerusalem, Israel) described the use of probabilistic models for the identification of regulatory networks and improved modelling of DNA-binding sites within proteins. He presented several applications in which graphical models have been helpful to integrate expression data with other genomic data to extract meaningful biological hypotheses and associate statistical confidence with them. These models have been used to identify new binding sites for transcription factors, interactions among transcription factors, co-regulated gene modules and to predict expression profiles for genes under various conditions. An example is the identification of a respiration module in *S. cerevisiae*, in which transcription factor PHD1 activates its target genes, including the gene for the transcription factor HAP4, and this factor then activates secondary target genes such as *COX4, COX6* and *ATP17* (Segal *et al*, 2003). T. Schlitt (Cambridge, UK) compared experimental data sets for *S. cerevisiae* using a

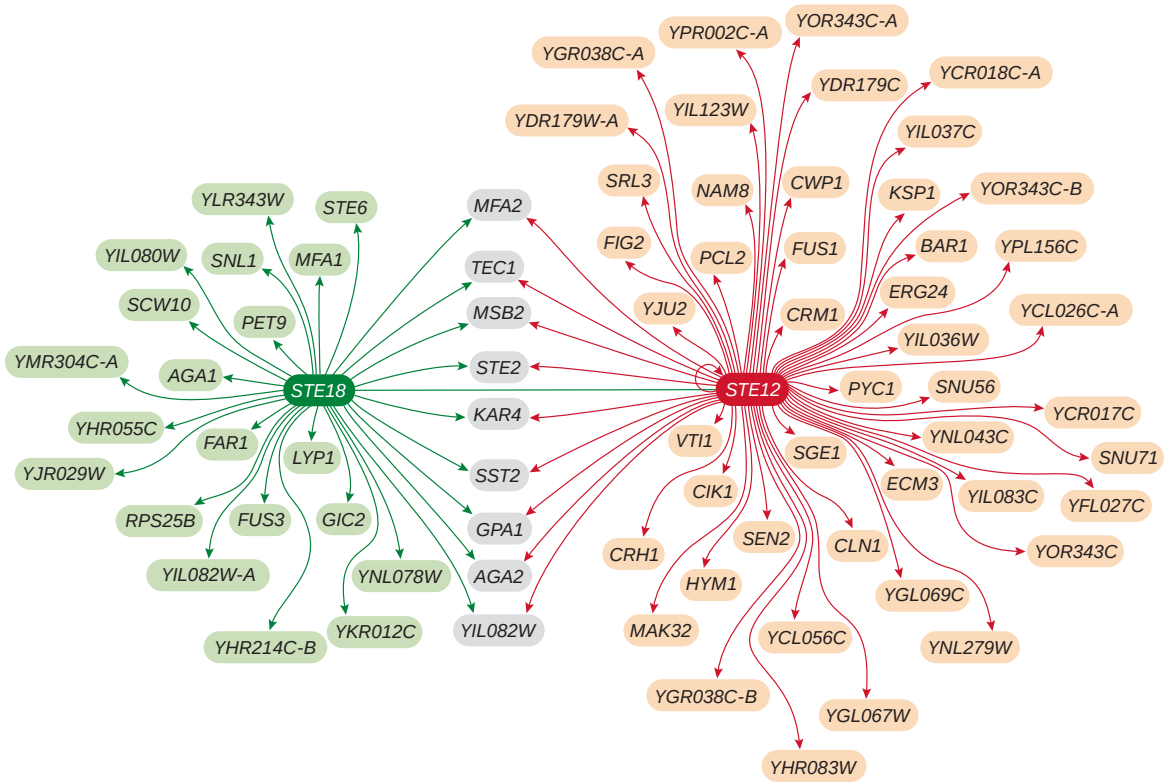

Fig 2 | Functional relationships between genes can be identified using a graph-based approach. A functional relationship between the two genes *STE12* and *STE18* is indicated by the small but significant overlap (grey nodes) between the yeast genes that have a binding site for the transcription factor STE12 in their promoter, and expression changes observed in the *Δste18*-deletion mutant. Here, genes are represented as nodes and are connected to *STE18* (dark green node) if their expression differs between the *Δste18* mutant and the wild-type strain (green connections), or to *STE12* (red node) if STE12 binds to their putative promoters (red connections). Both *STE12* and *STE18* are involved in the pheromone response in yeast.

graph-based approach, in which arcs (A→B) are used to represent information such as "transcription factor A regulates gene B" (Fig 2). The comparison includes data sets from chromatin-immunoprecipitation experiments, computational analyses of transcription-factor binding sites and microarray experiments on single-gene-deletion mutants (Schlitt *et al*, 2003). It is possible to predict functional relationships between genes by comparing gene neighbourhoods in these graphs (Fig 2). The relationships that were identified correspond to known protein–protein interactions and/or their co-occurrence in abstracts of scientific articles. F. Falciani (Birmingham, UK) used relevance networks to model the interaction between tumour cells and normal cells in prostate cancer. Crosstalk is thought to influence many important aspects of tumour biology. Falciani has developed a strategy to identify new genes that are involved in this crosstalk on the basis of gene-expression profiling and statistical modelling. The analysis revealed several genes, such as the repulsive factor *Slit-2*, that have a potential role in tumour growth and metastasis formation, and these have been verified experimentally.

Conclusion

During this workshop, numerous topics were discussed, ranging from data preprocessing to machine-learning approaches. Many challenges are still being faced with regard to the proper annotation of both the material-processing steps and the genes themselves. New insights are also being gained with respect to data normalization and preprocessing, in which methods to deal with changes in global mRNA expression and ratio statistics allow gene-expression changes to be measured more accurately. The talk dealing with *cis*-regulatory module detection showed that more sophisticated methods are available for this task, which are able to address large non-compact genomes. During the network session, it became clear that many challenges still exist in this area. Important aspects of how to deal with large data sets, different data qualities and different types of data are actively being explored. Progress has already been made in many of these areas and should continue in the future, with new developments and technologies becoming available. From this workshop, it was clear that the interplay between different areas of expertise will have a crucial role in advancing our understanding of biological processes.

ACKNOWLEDGEMENTS
The authors thank A. Brazma, M. Clark, F.C.P. Holstege and the speakers for their comments on the manuscript, as well as the organizers of the workshop. P.K. is supported by a grant from the Netherlands Organisation for Scientific Research (NWO; grant no. 05050205).

REFERENCES

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B (2003) Computational detection of *cis*-regulatory modules. *Bioinformatics* **19** (Suppl. II): 5–14
Al-Shahrour F, Díaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20:** 578–580
Ashburner M *et al* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29
Blaschke C, Oliveros JC, Valencia A (2001) Mining functional information associated with expression arrays. *Funct Integr Genomics* **1:** 256–268
Brazma A *et al* (2001) Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet* **29:** 365–371
Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* **31:** 3461–3467
Huber W, Von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** (Suppl.): S96–S104
Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK (2003) Predicting gene ontology biological process from temporal gene expression patterns. *Genome Res* **13:** 965–979
Lawrence CE, Altschul SF, Bogouski MS, Liu JS, Neuwald AF, Wooten JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:** 208–214
Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19:** 368–375
Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A (2003) From gene networks to gene function. *Genome Res* **13:** 2568–2576
Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166–176
Tanay A, Shamir R (2003) in *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 2003)* (eds Vingron M, Istrail S, Pevzner P, Waterman M), 301–310. ACM, New York, USA. doi:10.1145/640075.640115
Valentini G (2003) An application of low bias bagged SVMs to the classification of heterogeneous malignant tissues. *Lect Notes Comput Sci* **2859:** 316–321
van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4:** 387–393

Thomas Schlitt

Patrick Kemmeren