Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2011, **7**, 2863–2871

www.rsc.org/molecularbiosystems

PAPER

Linking molecular feature space and disease terms for the immunosuppressive drug rapamycin

Andreas Bernthaler,[a] Konrad Mönks,[b] Irmgard Mühlberger,[b] Bernd Mayer,[b] Paul Perco[b] and Rainer Oberbauer*[ac]

Received 19th May 2011, Accepted 29th June 2011
DOI: 10.1039/c1mb05187c

Next to development of novel drugs also drug repositioning appears promising for tackling unmet clinical needs. Here Omics provided the ground for novel analysis strategies for linking drug and disease by integrating profiles on the molecular as well as the clinical data level. We developed a workflow for linking drugs and diseases for identifying repositioning options, and exemplify the procedure for the immunosuppressive drug rapamycin. Our strategy rests on delineating a drug-specific molecular profile by combining Omics data reflecting the drug's impact on the cellular status as well as drug-associated molecular features extracted from the scientific literature. For rapamycin the respective profile held 905 unique molecular features reflecting defined molecular processes as identified by molecular pathway and process enrichment analysis. Literature mining identified 419 diseases significantly associated with this rapamycin molecular feature list, and transforming the significance of gene-disease associations into a continuous score allowed us to compute ROC and precision-recall for comparing this disease list with diseases already undergoing clinical trials utilizing rapamycin. The AUC of this assignment was computed as 0.84, indicating excellent recovery of relevant disease terms solely based on the drug molecular feature profile. We verified relevant indications by comparing molecular feature sets characteristic for the identified diseases to the drug molecular feature profile, demonstrating highly significant overlaps. The presented workflow allowed positive identification of diseases associated with rapamycin utilizing the drug-specific molecular feature profile, and may be well applicable to other drugs of interest.

Introduction

Drug development shows a thrilling timeline when thriving for novel drugs, and certainly significant further efforts are needed for tackling unmet clinical needs.[1–4] However, next to identification of novel drugs also repositioning of given drugs may be of value for combating yet unmet diseases.[5–8]

Over the last few years a number of approaches have been aimed at generating further knowledge about drug mode of action and the association to diseases emerged by systematically analyzing Omics data, clinical phenotypes, and drug effects in the context of biological networks. In a recent review Spiro *et al*. provided a summary on network representations for interlinking information from the patient phenotype, disease, therapy, drugs, and their molecular targets.[9] Butte *et al*. used transcriptomics data from the Gene Expression Omnibus database[10,11] for building a network of relations between transcriptomics features and phenotype characteristics.[12] Goh *et al*.[13] linked phenotypic states to the molecular level by exploiting known disease–gene associations from OMIM.[14] In this network the nodes represented diseases and edges represented the number of genes related to both diseases. Based on these data a sub-network of metabolic diseases was delineated, linking diseases *via* enzymatic reactions as stored in KEGG[15] and the BiGG database.[16] This approach demonstrated that strongly connected diseases exhibited several specific properties such as higher correlations in the reaction flux rate and expression of respective enzymes. These results indicated an increased likelihood for joint occurrence of diseases that exhibit connections on the level of metabolism.[17] Correspondingly, Hidalgo *et al*. showed that patients are more prone to develop comorbidities which—on the level of a network representation of affected molecular features—are close to a disease they are diagnosed with in the first place.[18]

The connectivity map as another systematic approach aims at generating drug-target networks utilizing gene expression data. In this approach transcriptomics data of human cancer cell lines were obtained reflecting the perturbation induced by a drug.[19] These profiles can now be utilized as a reference for

[a] *Krankenhaus der Elisabethinen Linz, Internal Medicine III, Department of Nephrology, Fadingerstrasse 1, 4010 Linz, Austria. Fax: +43 732 7676 4306; Tel: +43 732 7676 4300*
[b] *Emergentec Biodevelopment GmbH, Gersthofer Straße 29-31, 1180 Vienna, Austria*
[c] *Medical University of Vienna, Department of Internal Medicine III, Waehringer Guertel 18-20, 1090 Vienna, Austria. E-mail: rainer.oberbauer@meduniwien.ac.at*

matching disease associated transcriptomics data and the profiles provided in the connectivity map, providing drugs whose mode of action relates to the disease profile on the transcript profile level. These approaches serve as examples for a number of large-scale mappings of drugs and diseases in a network context.[20–25]

In the realm of drugs, induced changes on a cellular transcriptome and proteome level, and available data on diseases enable us to propose a workflow aimed at relating drugs to diseases based on joint molecular feature spaces. We exemplify this approach for the immunosuppressive and antiproliferative drug rapamycin, also known as sirolimus.[26] Rapamycin inhibits mTOR1 which plays a vital role in a variety of key cellular functions such as cell growth and nutrient utilization. The drug has been associated to a spectrum of disease terms such as graft rejection, autoimmune disorders, cardiovascular diseases, and metabolic disorders,[26] as well as a broad variety of renal diseases[27] and neoplasms.[26,28,29]

Our approach centrally rests on delineation of a drug-specific molecular feature set reflecting the impact of the drug on the cellular level beyond the primary target, subsequently using this drug-specific profile for identification of diseases also exhibiting perturbation in this profile. Potential application areas of rapamycin are presented and discussed in the context of recent literature and clinical trial data. Highly ranked diseases are finally analyzed in more detail on the level of molecular features in comparison to the rapamycin molecular profile, further strengthening the validity of our approach.

Results

Rapamycin molecular feature sets

Six individual molecular feature sets associated with rapamycin mode of action and more generally impact on cellular processes were extracted and linked to diseases, as listed in Table 1.

Differentially regulated features for PBMCs (PBMC-profile, 1407 differentially regulated probes mapping to 184 unique genes), differentially regulated features extracted from the connectivity map data (CMAP-profile, 304 differentially regulated probes matching to 266 unique genes), the data set derived from KEGG and expert reviews (63 unique genes), and profiles from scientific literature search utilizing MeSH and Fable (in total 468 unique genes resting on partially overlapping 196 genes from MeSH-based literature search and 372 genes from Fable).

To gain insight into the homogeneity of the individual feature sets we analyzed their overlap on the level of gene symbols, as shown in Fig. 1. Next to direct feature overlap we analyzed the consensus between the data sets using the PANTHER classification system (Protein ANalysis THrough Evolutionary Relationships) for molecular pathways and processes.[30]

Table 1 Summary of rapamycin associated molecular features and diseases

Source	Short name	# Genes	# Diseases
Transcriptomics features	CMAP	266	1213
Transcriptomics features	PBMC	184	1635
Literature features	AUTO	468	2779
Literature features	EXP	63	1457
Keyword search	TRIALS	—	425

Source and short name for data sets used, the number of associated molecular features, and the number of assigned diseases.

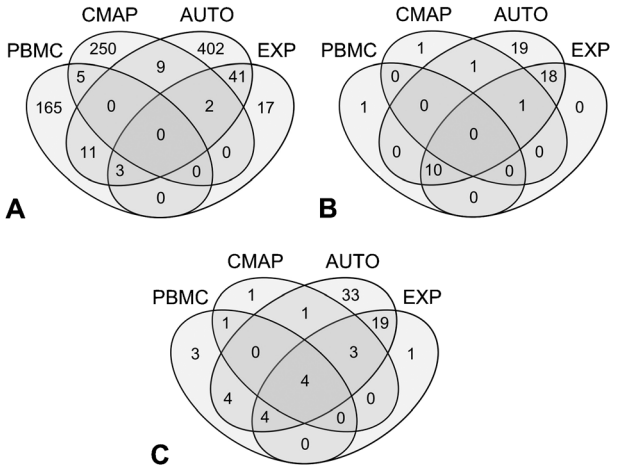

Fig. 1 Molecular feature sets linked to rapamycin as derived on the basis of PBMC expression (PBMC), the connectivity map (CMAP), automated literature search following two different extraction approaches (AUTO), and database information together with expert reviewing (EXP) are displayed on the level of direct feature overlap as well as pathway and process overlap: (A) direct feature overlap of the two transcriptomics, the consolidated literature, and database/expert data set, (B) overlap on the level of pathways, (C) overlap on the level of enriched biological processes.

The direct overlap of genes in the four profiles was found as minor (Fig. 1A), *i.e.* a significant fraction of genes is apparently specific for the respective source. Only five genes (BRAF, MAPK1, RHEB, VEGFA, EIF4B) occurred in at least three feature sets. This heterogeneity is particularly evident for the transcriptomics profiles, and considerably less pronounced when comparing the database/expert profile and the literature data set. On the level of PANTHER pathway enrichment (Fig. 1B) and process enrichment (Fig. 1C) a pronounced overlap is found for the given data sets. Significantly enriched pathways and processes including at least three molecular feature sets are listed in Table 2. The two feature profiles based on automatic and expert literature search were combined due to the high overlap of these sets resulting in a profile of 485 unique molecular features. This profile was used for all subsequent analyses.

Rapamycin associated clinical trial data

575 clinical trials holding rapamycin in the trial description were retrieved from ClinicalTrials.gov. These trials cover all clinical study phases, and focus primarily on three disease categories, namely cardiovascular diseases (CARDIO), neoplasms (NEOPLASM), and immune-related indications (IMMUNO, including 'transplant', 'immune', 'graft', 'graft-*versus*-host'), as displayed in Fig. 2. These three disease categories covered 501 of the in total 575 identified trials, leaving 74 trials to other diseases (OTHER). Of the rapamycin associated clinical trials 256 are in Phase I, 311 in Phase II, 145 in Phase III, and 201 in Phase IV.

Table 2 Significantly enriched PANTHER pathways and processes

(A) Pathway	EXP-profile	AUTO-profile	PBMC-profile	CMAP-profile
Insulin/IGF pathway–protein kinase B signaling cascade	7.78×10^{-28}	6.05×10^{-21}	2.88×10^{-02}	—
Ras Pathway	8.68×10^{-25}	4.33×10^{-31}	2.12×10^{-02}	—
T cell activation	3.26×10^{-21}	1.00×10^{-37}	7.20×10^{-03}	—
Interleukin signaling pathway	6.43×10^{-20}	3.73×10^{-38}	7.17×10^{-03}	—
p53 pathway	8.10×10^{-19}	1.33×10^{-20}	—	1.59×10^{-02}
Apoptosis signaling pathway	1.90×10^{-13}	1.44×10^{-34}	1.35×10^{-02}	—
Inflammation mediated by chemokine and cytokine signaling pathway	2.11×10^{-13}	3.08×10^{-32}	5.32×10^{-03}	—
B cell activation	5.68×10^{-12}	9.93×10^{-27}	2.34×10^{-02}	—
TGF-beta signaling pathway	1.12×10^{-03}	6.15×10^{-09}	2.31×10^{-02}	—
Interferon-gamma signaling pathway	3.79×10^{-03}	3.42×10^{-07}	1.98×10^{-02}	—
Parkinson's disease	3.91×10^{-02}	2.93×10^{-06}	3.86×10^{-02}	—
(B) Process				
Intracellular signaling cascade	2.43×10^{-18}	8.37×10^{-58}	4.12×10^{-02}	—
Cell cycle	1.61×10^{-13}	2.45×10^{-39}	—	4.20×10^{-03}
Protein metabolic process	1.55×10^{-10}	1.62×10^{-34}	6.73×10^{-03}	3.15×10^{-10}
Response to stress	4.22×10^{-08}	2.05×10^{-30}	1.30×10^{-03}	8.35×10^{-04}
Apoptosis	7.27×10^{-06}	1.59×10^{-48}	2.19×10^{-03}	—
Primary metabolic process	8.83×10^{-06}	1.08×10^{-21}	1.05×10^{-02}	8.56×10^{-09}
Mitosis	2.56×10^{-05}	6.10×10^{-13}	—	1.57×10^{-03}
Metabolic process	2.74×10^{-05}	7.00×10^{-21}	8.08×10^{-03}	1.99×10^{-08}
Phosphate metabolic process	9.66×10^{-05}	2.15×10^{-08}	8.50×10^{-03}	—
Meiosis	3.42×10^{-04}	3.04×10^{-09}	—	1.26×10^{-02}
Induction of apoptosis	5.20×10^{-03}	4.99×10^{-16}	1.76×10^{-02}	—

(A) Pathway and (B) process terms as provided in PANTHER, and respective enrichment expressed as *p*-values for the four data sets.

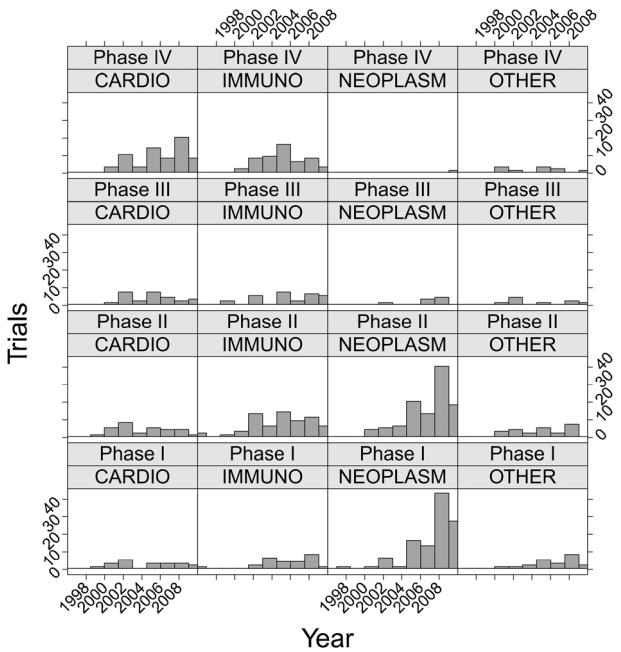

Fig. 2 Rapamycin associated clinical trials as registered in ClinicalTrials.gov, and the respective number of trials per year, indication (CARDIO, NEOPLASM, IMMUNO, OTHER), and clinical phase.

Rapamycin and disease associations

A combined disease score indicative for the evidence of a gene–disease link could be calculated for 2859 disease terms on the basis of the molecular feature lists (see Table 1), however most of the diseases show weak associations. Disease lists derived from transcriptomics feature lists showed a very specific response pattern despite the heterogeneity inherited in the transcriptomics datasets. For unraveling the specific association of diseases derived on the basis of the Omics lists we compared the distributions of disease scores of both disease lists derived from transcriptomics against disease scores derived from a random feature list of the same size. Distributions of disease scores deviated significantly (two sided, two sample Kolmogorov–Smirnov test) yielding significantly lower combined disease scores for the random list (PBMC: $p < 1.8 \times 10^{-7}$, CMAP: $p < 1.5 \times 10^{-9}$).

For further exploring the relevance of our transcriptomics list-associated diseases regarding rapamycin, we generated molecular feature profiles using automated literature search for four additional drugs, namely Diclofenac, Nifedipine, Diazepam, and Ciprofloxacin. Disease scores were obtained for these drugs using the same approach as for the rapamycin molecular feature profile, and Pearson's correlation coefficients between the disease scores were compared. Disease scores of our combined transcriptomics profiles exhibited significantly higher correlations ($r = 0.7$) to the expert-curated rapamycin data set than correlations of disease scores of the four other drugs ($\mu = 0.12$, $\sigma = 0.2$), further strengthening mTOR specific response for the transcriptomics data on the level of disease scores.

The rapamycin specific combined disease score assigned to each disease term was used for identifying the recovery of disease terms effectively under study in clinical trials for the drug rapamycin. From the MeSH hierarchy we obtained 2859 disease terms with 425 diseases already mentioned in clinical trials on rapamycin representing our set of positive instances (disease terms with at least one clinical trial assigned), and 2434 negative instances (disease terms with no clinical trial assigned). Generation of the ROC curve was based on the ranked list of diseases based on our calculated disease score with MeSH disease terms already investigated in a clinical trial

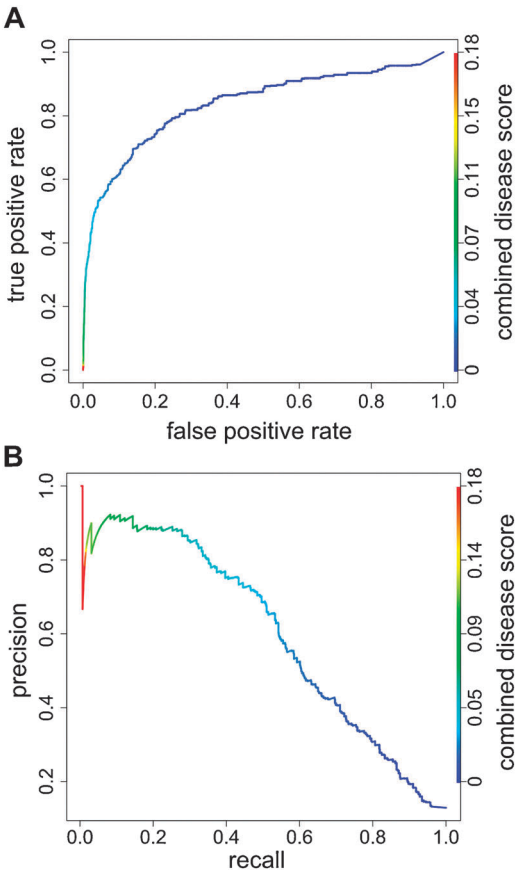

Fig. 3 (A) Receiver operating characteristic and (B) precision-recall curve of the combined disease score as predictor for identifying rapamycin-associated disease terms investigated in clinical trials.

considered as hits and all other disease terms as non-hits. Fig. 3 displays the resulting ROC curve with an AUC of 0.84 (Fig. 3A) and the respective precision-recall curve (Fig. 3B). For further analysis we only considered disease associations with a significantly enriched combined disease score (FPR $<$ 5%, combined disease score $>$ 0.026), resulting in 419 rapamycin-associated disease MeSH-terms (PPV = 62%, NPV = 93%).

250 out of the 419 disease terms delineated on the basis of the rapamycin molecular feature set were already mentioned in at least one clinical trial, whereas for the remaining 169 disease terms no clinical trials were identified. MeSH terms found for these 169 disease terms included generic terms such as 'Drug Toxicity', 'Animal Diseases', 'Disease', and were, next to terms potentially associated with non-therapeutic effects of using immunosuppressive medication (Bacterial Infections and Mycoses, Virus Diseases, Parasitic Diseases, Postoperative Complications, Pregnancy Complications), removed from the disease term list as well as the very general disease terms of level one and two of the hierarchy.

According to the MeSH hierarchy and subsequent expert classification the remaining list of potentially novel indications contained 115 disease MeSH terms related to mainly six disease areas, being cancer (Neoplasms, Hyperplasia, Polyps, Growth Disorders), nervous system-related diseases (Nervous System Diseases), inflammatory system related diseases (Inflammation, Systemic Inflammatory Response), metabolic diseases (Nutritional and Metabolic Diseases, Body Weight, Overweight), cardiovascular diseases (Cardiovascular Diseases), and diseases of the respiratory system (Respiratory Tract Diseases, Pulmonary Diseases, Asthma, Bronchitis), representing 28, 20, 12, 11, 9, and 8 out of the 115 disease MeSH-terms, respectively. These six classes of diseases covered 84 unique disease MeSH-terms and represented 71.3% of the identified diseases. When further including cancer related disease MeSH-terms (such as chromosome aberrations and cell growth defects) into the set of neoplasm related disease MeSH-terms, those disease areas covered 78.2% of the final set of diseases, leaving only 25 disease MeSH-terms to other diseases. A representative scheme regarding the coverage of the MeSH ontology with the identified disease terms is provided in Fig. 4A.

Feature profiles of rapamycin associated diseases

We validated feature profiles of highly ranked diseases from the area of neurodegenerative diseases (Parkinson's disease, Alzheimer's disease, dementia), neoplasms (B-cell lymphoma), and one disease affecting both disease areas (Neuroblastoma) by comparing disease-specific features characterizing these diseases to the rapamycin molecular feature profile.

For a graphical representation of our results we extracted the sub-graph of the disease MeSH-term 'B-Cell Lymphoma', member of the disease terms 'Immunoproliferative Disorders', 'Lymphatic Diseases', and 'Neoplasms by Histologic Type', from the scheme presented in Fig. 4A in further detail (Fig. 4B). Multiple studies investigating B-cell lymphoma and rapamycin are already registered rendering it a good instance for exemplification. The lymphoma subgraph further splits into three disease MeSH-terms being identified on the basis of the rapamycin-derived molecular feature sets and also being identified in the clinical trial data set ('Lymphoma, B-Cell', 'Lymphoma, Large, B-Cell, Diffuse', 'Burkitt's Lymphoma'). Two further terms were neither identified on the basis of the molecular feature sets nor from mining of clinical trials descriptions ('Lymphoma, AIDS-Related', 'Lymphoma, Primary Effusion'), and one sub-term was mentioned as being investigated in a clinical trial but was not identified on the basis of the molecular feature set ('Lymphoma, B-Cell, Marginal Zone'). We obtained transcriptomics data characterizing patients suffering from diffuse large B-cell lymphoma[41] and compared this feature set to the combined rapamycin-associated feature sets. 175 molecular features were identified as members of both, the B-Cell transcriptomics profile (consisting of 1416 unique genes) and the set of genes derived as rapamycin affected (comprised of in total 905 unique genes), representing a highly significant overlap (Fisher's exact test, $p < 2.2 \times 10^{-16}$). This close resemblance of molecular features associated with rapamycin, which in turn significantly overlap with molecular features being affected in the course of B-cell lymphoma, indirectly links the drug with this disease term, also becoming evident by two recently registered trials in Phase I/II (NCT01075321, NCT00918333) investigating the role of rapamycin in diffuse large B-cell lymphoma.

The second feature profile derived from transcriptomics data of patients suffering from Parkinson's disease yielded

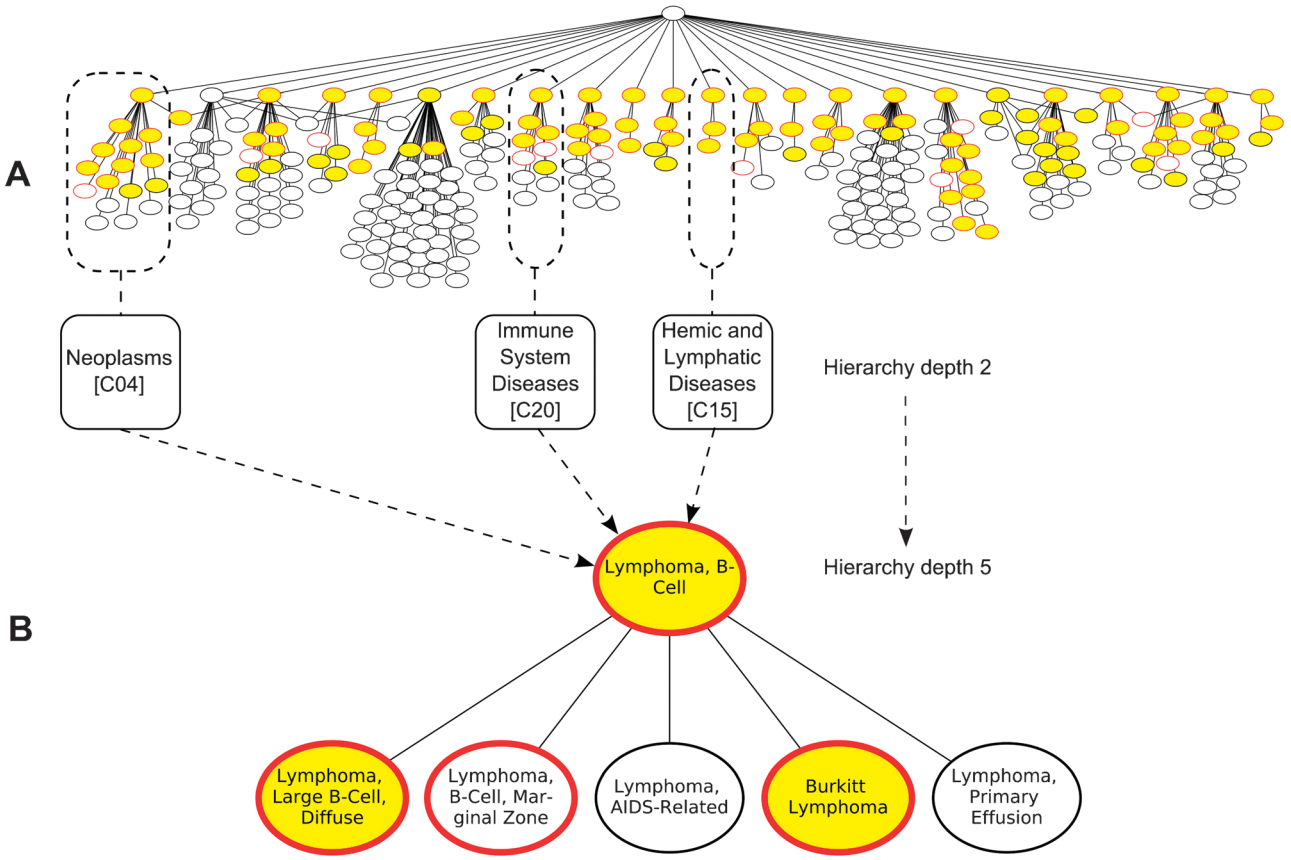

Fig. 4 Disease enrichment and rapamycin associated clinical trials. (A) MeSH-term hierarchy levels 1–3 of the category 'Disease [C]', and disease terms being identified either on the basis of the molecular feature lists or found in clinical trials given in color. (B) Excerpt of the hierarchy for the disease MeSH-term 'Lymphoma, B-Cell' (MeSH hierarchy level five) and the related sub-terms. MeSH-terms displayed in yellow were identified as being affected by rapamycin on the level of molecular features, and the disease terms encircled in red are represented at least in one clinical trial.

279 features significantly associated to the disease, and for this data set also a significant overlap (26 features) to the rapamycin molecular feature profile was identified ($p = 2.27 \times 10^{-6}$). We further retrieved feature profiles from the scientific literature for the MeSH terms 'Neuroblastoma', 'Alzheimer's disease', and 'Dementia' for validating other significant diseases as derived from our rapamycin molecular feature profile. According to our approach we obtained 206, 165, and 188 significantly associated features for these diseases, and 23, 21, and 21 overlapping features to our rapamycin molecular feature profile were found to be significant for all three disease terms ('Neuroblastoma': $p = 3.85 \times 10^{-7}$, 'Alzheimer's disease': $p = 1.35 \times 10^{-7}$, 'Dementia': $p = 1.21 \times 10^{-6}$).

Discussion

Two distinct data sources, namely Omics and scientific literature, were used for deriving a rapamycin-associated molecular profile. Regarding Omics data two specific transcriptomics data sets were used, one reflecting the impact of rapamycin on PBMCs, the other monitoring drug-induced changes on the level of three cancer cell lines as derived from the connectivity map. Certainly both data sets have limitations, including principal heterogeneity of PBMCs even within the very same donor, and *per se* altered biology of immortalized cell lines. The heterogeneity for the given transcriptomics data also became clear when comparing the rapamycin-associated features presented in the two data sets holding 184 and 266 differentially regulated genes, respectively.

The direct overlap of the two literature profiles exhibited higher concordance, with the first set obtained by merging two automated literature mining approaches resulting in 468 genes, and the second set obtained by merging rapamycin/mTOR specific signatures from expert curation yielding 63 genes. This increased overlap is not surprising, as both database knowledge in KEGG as well as expert reviewing naturally resemble information also published in the scientific literature. This automated and expert driven literature profiles were subsequently merged resulting in a combined set of 485 features in order to avoid a bias towards findings from the scientific literature in our analysis workflow.

A number of meta-studies on Omics data have been published describing a weak feature overlap even for more homogeneous studies, but this apparent complexity vanishes to some extent when traversing the analysis to the level of molecular processes and pathways. This finding reflects the assumption that functional units are consistently found *via* Omics studies, but individual features reflect specifically the samples and conditions used for analysis. In the PANTHER classification system eleven pathways were found to be significantly enriched by at least three feature data sets. Nine out of the eleven pathways are related to metabolism

(Insulin/IGF pathway–protein kinase B signaling cascade), cell growth and fate (Apoptosis signaling pathway, TGF-beta signaling pathway, p53 pathway), and immune system response such as cell communication (Interleukin signaling pathway, inflammation mediated by chemokine and cytokine signaling pathway, Interferon-gamma signaling pathway) and leukocyte activation (T cell activation, B cell activation). The remaining two pathways were the Ras-pathway and the pathway Parkinson's disease.

Ten pathways identified as significantly affected for the PBMC data set and one pathway from the CMAP data set were also identified when using the literature data sets for pathway enrichment analysis. For the literature profiles as such there was a 100% overlap of significantly enriched pathways compared to an overlap of 29% considering the explicit molecular feature sets. Biological processes significantly associated to at least three data sets primarily covered metabolic- and cell cycle-related processes, recovering eight PBMC processes and seven CMAP processes by the literature data sets, with four processes recovered in both transcriptomics profiles simultaneously indicating that the data sets reach coherence on the level of processes and pathways.

First-in-man studies of rapamycin have been documented from the year 1996 onwards, initially investigating patients needing immunosuppressive therapy after organ transplantation. Because of interference with T-cell activation, rapamycin is a potent medication for hindering allograft rejection, as well as in the context of graft-*versus*-host disease.[26] Next to interfering with immune activation rapamycin has been applied for patients suffering from coronary artery diseases using rapamycin coated stents to prevent restenosis after balloon angioplasty, and neoplasm, grounded on the drug's impact on cell cycle related biological processes.[26] Quite recently mTOR has been associated to cell growth defects with cancer as the most representative disease instance,[29] and this finding is reflected by the increasing number of trials investigating the role of rapamycin in this area especially within the past three years.

Retrieving rapamycin-associated clinical trials from clinicaltrials.gov resulted essentially in three disease categories, namely neoplasms, diseases affecting the immune system, and diseases affecting the cardiovascular system. On the basis of our workflow six main disease groups were identified, namely neoplasms, diseases affecting the nervous, the immune, the metabolic, the cardiovascular, and the respiratory system, with these six groups accounting for 78.2% of all disease terms identified. The group of cardiovascular diseases and angiogenesis related effects of rapamycin as well as the association to the immune system are well discussed in the scientific literature.[26] Furthermore, rapamycin responds to insulin like growth factors *via* the PI3K pathway, and as nutrition heavily regulates mTOR activity a link to the metabolism related diseases can be assumed,[26] with disease terms such as obesity, overnutrition, overweight, and several other metabolism related appearances being identified. Recently, Harrison *et al*. showed that rapamycin extended the life span of mice and the authors suspected intervening effects of rapamycin related to retarding mechanisms of aging.[31] The cancer related disease MeSH-terms identified ('B-Cell Lymphoma', 'Diffuse Large B-Cell Lymphoma' and 'Burkitt's Lymphoma') were discussed in the scientific literature[29] as being related to the PI3K-AKT-mTOR pathway and validation of molecular features characteristic for B-cell lymphoma was recovered in the drug molecular feature profile of rapamycin exhibiting highly significant overlap. Recently the effect of rapamycin on nervous system related functions is drawing attention. Hu *et al*. identified rapamycin related to Huntington disease[24] which is supported by Sarkar *et al*.,[32] and Pan *et al*. showed that rapamycin may enhance autophagy in response to nerve cell degeneration resulting in a neuroprotective effect.[33] Malagelada *et al*. found neurons in animal toxin models of Parkinson's disease as being protected from death upon rapamycin intervention,[34] and feature profiles characteristic for Alzheimer's disease, Parkinson's disease, and Dementia were recovered in the rapamycin specific feature set with high significance. Furthermore, rapamycin was associated to respiratory tract diseases in the recent literature.[35]

Interestingly, Neuroblastoma—a disease falling into two identified disease categories of interest, namely nervous system related diseases and neoplasms—was identified as the top hit of our analysis, and subsequently, molecular features associated to this disease were recovered in the rapamycin molecular feature profile with significant overlap.

Experimental

Drug repositioning workflow

Our workflow consists of four data processing and analysis steps as schematically depicted in Fig. 5.

The first step was the identification of features identified as annotated in the context of rapamycin utilizing transcriptomics profiles of rapamycin treated samples in a case/control study design, as well as features associated to the drug according to the scientific literature. Based on this set of molecular features reflecting the impact of rapamycin on the cellular feature level (rapamycin molecular feature profile) we performed a literature analysis for delineating links between these given molecular features and all diseases discussed specifically in the context of these particular features. MeSH-terms for disease association were extracted for PubMed articles holding at least one member of the rapamycin molecular feature list. This procedure resulted in a ranked list of diseases, where the rank was defined by the frequency of a particular disease being identified on the basis of the molecular feature list. In step 2 we retrieved clinical trial information available for rapamycin (including its synonyms) and extracted diseases from the trial descriptions, thus obtaining a list of diseases effectively studied in a clinical setting in the context of rapamycin. In step 3 we identified disease MeSH-terms being frequently found on the basis of the molecular feature list generated in step 1 but not being covered in clinical trials, resulting in potentially novel indications. In the validation step we used highly ranked disease terms, namely B-cell lymphoma, Neuroblastoma, Alzheimer's disease, Parkinson's disease, and dementia, for discussing transcriptomics and literature data available in the light of the molecular feature list found as being associated with rapamycin.

Rapamycin associated molecular features

A first set of rapamycin associated molecular features was derived from representative transcriptomics data sets using the

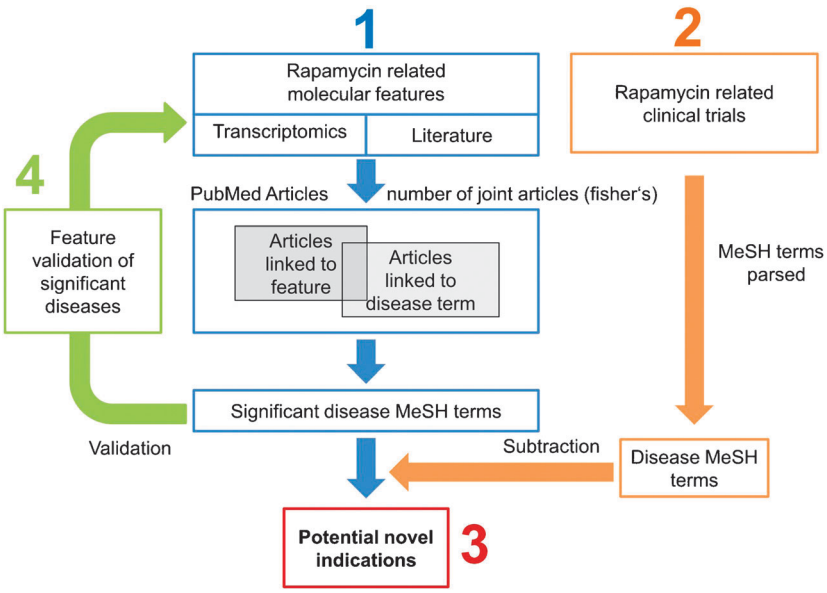

Fig. 5 Schematic overview of the analysis procedure: (1) retrieval of molecular features related to rapamycin based on Omics profiles and scientific literature, (2) retrieval of disease MeSH-terms associated to rapamycin as found in clinical trial descriptions, (3) comparison of disease lists obtained in (1) and (2). For validation purposes (4) exemplary retrieval of features of highly ranked diseases for comparison with the feature list generated in (1) is included.

Table 3 Transcriptomics data used for extraction of rapamycin-associated molecular features

Reference	Profile	Samples	Specimen characteristics	Platform
GSE12187[37]	PBMC-profile	18 Rapamycin treated	Peripheral blood mononuclear cells	HG-U133A + 2
GSE8507[36]	PBMC-profile	17 Control samples	Peripheral blood mononuclear cells	HG-U133A + 2
GSE5258[19]	CMAP-profile	14 Rapamycin treated/14 controls (34/63 including technical replicates)	Cell lines MCL-7, HL60, and PC3	HT_HG-U133A

Listed are reference (GEO accession number), profile type, number of samples in case and control, cell type, and platform used for expression profiling.

drug in cell line settings as listed in Table 3. We selected two studies from the Gene Expression Omnibus (GEO) database[10] for analysis, comparing differentially expressed features from rapamycin treated peripheral blood mononuclear cell (PBMC) samples to non-treated PBMCs.[36,37] Treated PBMC samples originated from patients undergoing immunosuppressive therapy (combination of cyclosporine and rapamycin) after renal transplantation. The other set contained untreated PBMC samples from healthy donors, allowing a comparison of 18 rapamycin treated samples to 17 controls, accessed and downloaded from GEO (April, 2010). Affymetrix CEL files were RMA normalized and Significance Analysis of Microarray (SAM) was used for identifying differentially regulated features ($d = 2.8$, 300 permutations, 0% FPR, fold change > 3).

Next to the PBMC profiles expression data as recorded in the connectivity map (build 2.0) were used.[19] The connectivity map documents the response in gene expression to drug treatment containing about 7000 profiles based on 1309 compounds. As of April 2010 the connectivity map contained 97 expression profiles using the Affymetrix HT-HG-U133A platform characterizing rapamycin treated cell lines. Consolidation of technical replicates led to in total 14 samples for case and control each, recorded for the cancer cell lines MFC7, PC3, and HL60. As for the PBMC samples the raw Affymetrix data were RMA normalized subsequently identifying differentially regulated features applying a paired *t*-test with adjusted Bonferroni correction for multiple testing ($p < 0.05$). Analysis was performed using the affy and limma packages in R/BioConductor.

Complementary to identifying rapamycin-associated features on the basis of Omics studies automated literature screening for feature identification was performed. The first approach relied on screening for features enriched in the set of rapamycin-associated articles as provided in PubMed.[11] The respective set of papers was extracted from PubMed using the Mesh-term 'rapamycin' (including the synonyms 'Sirolimus', 'AY 22-989', and 'Rapamune'), the mapping of genes to articles was retrieved from the NCBI gene-2-pubmed file (April 2010, ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz). Based on the association of papers and genes we determined the frequency of occurrences for each gene within the set of rapamycin-associated publications. Taking the background frequency of each gene in all PubMed listed articles into account a Fisher's exact test was applied for identifying genes being significantly associated with the drug (MeSH-approach).

In a second approach we used FABLE (Fast Automated Biomedical Literature Extraction),[38–40] and genes being cited

at least ten times in the context of the drug were further considered. Finally, features from the MeSH-approach and the Fable-based search were combined into one feature profile. Additionally genes were extracted from a recent review discussing mTOR molecular mode of action,[26] and further extended by genes represented in the mTOR signaling pathway (hsa04150) of KEGG.[15] For characterizing features from automated and expert search, pathway enrichment analysis was performed separately for these profiles (AUTO-, EXP-profile). According to the large overlap of these profiles on the feature level we assumed those two sets are not independent. Hence, subsequent analysis of disease associations was performed using the combined set of automated and expert based feature profiles.

As a negative control, features associated to drugs not related to rapamycin were obtained for Diclofenac, Nifedipine, Diazepam, and Ciprofloxacin using our MeSH-approach as described above. Only features with an association exhibiting $p < 0.05$ were included in the unrelated drug feature profiles.

For validating diseases significantly associated to rapamycin we obtained molecular feature sets for Neuroblastoma, Alzheimer's disease, dementia, Parkinson's disease, and diffuse large B-Cell lymphoma. For these diseases an explicit validation was performed. For B-cell lymphoma we used a feature list of transcriptomics data provided by ONCOMINE[41] as consolidated in,[42] and compared this molecular feature list to the rapamycin molecular feature set. Associated features for Parkinson's disease were extracted from transcriptomics data comparing individuals that developed Parkinson's disease and healthy controls.[43] Processed data are accessible through the curated dataset browser of the NCBI GEO database[10] (GEO accession GDS6613, $p < 0.001$). Features being associated with Neuroblastoma, Alzheimer's disease, and dementia were obtained using our MeSH-based feature extraction method. Genes with significant p-values ($p < 0.05$) were included in these disease specific feature profiles.

Rapamycin associated diseases

To obtain rapamycin-associated diseases being investigated in clinical trials we queried ClinicalTrials.gov (September 2010, www.clinicaltrials.gov) using the keyword rapamycin (and expanding the search term to 'Sirolimus', 'AY 22-989', and 'Rapamune'),[44] subsequently retrieved their associated MeSH-terms, and extracted the terms holding disease information from the MeSH category 'Disease [C]'.

For retrieving disease terms associated to the rapamycin molecular features we identified PubMed articles linked to these features and extracted the disease MeSH-terms of these articles. For this we identified the set of articles being linked to each of the molecular features provided in the rapamycin molecular feature profile using the NCBI gene-2-pubmed mapping (accessed in April 2010, available at ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz). For each disease MeSH term we then determined the set of articles associated to this term with respect to the entire PubMed set of articles using NCBI Entrez Utilities.[11] This allowed the application of a Fisher's exact test on the overlapping set of articles and subsequent association of each gene in the feature profiles to its disease MeSH terms with a p-value. We derived a set of disease MeSH terms for each of the three feature profiles, each MeSH term being described by a set of gene associations in the form of p-values. For each MeSH term the p-values in the respective sets were transformed into a rank

$$\text{rank} = \frac{\sum_{i=1}^{n} 1 - p_i}{n}$$

where n is the number of genes that could be associated to a MeSH term in the respective feature profile, and p_i is the p-value obtained by the Fisher's exact test of each gene. We used the arithmetic mean of the *rank* of each disease MeSH term to combine the disease sets derived from the three feature profiles for ranking the significance of disease association leading to a combined disease score.

Respective diseases for the drugs Diclofenac, Nifedipine, Diazepam, and Ciprofloxacin were obtained following the same procedure.

Based on the combined disease score assigned to the disease MeSH terms we measured performance of the combined disease score using the rocr package of the statistical software R. We calculated the ROC curve, and complementary, due to imbalanced distribution of classes (425 positive/2434 negative instances), the precision-recall curve. The combined disease score was validated using information about whether a disease already underwent clinical investigation according to the data reported in www.clinicalTrials.gov. Diseases registered at least once in a clinical study constituted a hit (representing our set of positive instances) and disease MeSH terms occurring to have no clinical study assigned constituted a miss (negative instances).

Conclusions

Using the level of significance of a disease being associated to the molecular feature list as scoring criterion allowed us to compare molecular feature set-derived disease terms and terms retrieved from clinical trial descriptors on the level of a ROC curve, *i.e.* computing sensitivity and specificity with respect to missed disease terms but also novel disease terms. The AUC value of 0.84 and the precision recall curve clearly indicated an excellent recovery of disease terms already considered in clinical trials, but also provided novel disease terms (in the context of the ROC curve to be seen as false positives). Out of 419 disease terms significantly associated to rapamycin as delineated on the basis of the rapamycin molecular feature profile 115 novel disease terms were detected.

We exemplarily depicted Neuroblastoma, Alzheimer's disease, Parkinson's disease, dementia, and B-cell lymphoma as a relevant indication in this context for testing the relation of these disease terms on the level of affected molecular features. Consensus lists of features from literature and differential expression data of these diseases showed significant overlap with the molecular feature profile assigned to the drug. Relating drug-specific Omics profiles and drug related features derived from the scientific literature to diseases promises correct assignment of relevant diseases. The example on rapamycin presented in this work provided positive validation of this workflow, as demonstrated by the significant

overlap of disease terms derived on the basis of the molecular feature profile and the spectrum of clinical trials already investigating the therapeutic effect of this drug, but also by the high overlap of molecular features found affected for a particular novel indication and the drug-specific molecular profile initially used for screening novel indications.

Acknowledgements

This work was partly supported by Wyeth Lederle Pharma GmbH, Austria.

Notes and references

1 C. P. Adams and V. V. Brantner, *Health Aff.*, 2006, **25**, 420–428.
2 J. DiMasi, *J. Health Econ.*, 2003, **22**, 151–185.
3 J. Drews, *Science*, 2000, **287**, 1960–1964.
4 L. J. Gershell and J. H. Atkins, *Nat. Rev. Drug Discovery*, 2003, **2**, 321–327.
5 J. Drews, *Nat. Rev. Drug Discovery*, 2006, **5**, 975.
6 T. T. Ashburn and K. B. Thor, *Nat. Rev. Drug Discovery*, 2004, **3**, 673–683.
7 B. Munos, *Nat. Rev. Drug Discovery*, 2009, **8**, 959–968.
8 S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nat. Rev. Drug Discovery*, 2010, **9**, 203–214.
9 Z. Spiro, I. A. Kovacs and P. Csermely, *J. Biol.*, 2008, **7**, 20.
10 T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, *Nucleic Acids Res.*, 2007, **35**, D760–D765.
11 E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. John Wilbur, E. Yaschenko and J. Ye, *Nucleic Acids Res.*, 2010, **38**, D5–D16.
12 A. J. Butte and I. S. Kohane, *Nat. Biotechnol.*, 2006, **24**, 55–62.
13 K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabási, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–8690.
14 V. A. McKusick, *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*, Johns Hopkins University Press, Baltimore, 12th edn, 1998.
15 M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, *Nucleic Acids Res.*, 2002, **30**, 42–46.
16 J. Schellenberger, J. O. Park, T. M. Conrad and B. Ø. Palsson, *BMC Bioinf.*, 2010, **11**, 213.
17 D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai and A.-L. Barabási, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 9880–9885.
18 C. A. Hidalgo, N. Blumm, A.-L. Barabási and N. A. Christakis, *PLoS Comput. Biol.*, 2009, **5**, e1000353.
19 J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander and T. R. Golub, *Science*, 2006, **313**, 1929–1935.
20 F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi and D. di Bernardo, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, 107.
21 S. Zhao and S. Li, *PLoS One*, 2010, **5**, e11764.
22 M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork, *Science*, 2008, **321**, 263–266.
23 M. a Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási and M. Vidal, *Nat. Biotechnol.*, 2007, **25**, 1119–1126.
24 G. Hu and P. Agarwal, *PLoS One*, 2009, **4**, e6536.
25 Y. Y. Li, J. An and S. J. M. Jones, *Genome informatics. International Conference on Genome Informatics, Pacifico Yokohama*, 2006, vol. 17, pp. 239–247.
26 S. Wullschleger, R. Loewith and M. N. Hall, *Cell*, 2006, **124**, 471–484.
27 W. Lieberthal and J. S. Levine, *Journal J. Am. Soc. Nephrol.*, 2009, **20**, 2493–2502.
28 D. a Guertin and D. M. Sabatini, *Trends Mol. Med.*, 2005, **11**, 353–361.
29 D. a Guertin and D. M. Sabatini, *Cancer Cell*, 2007, **12**, 9–22.
30 H. Mi, N. Guo, A. Kejariwal and P. D. Thomas, *Nucleic Acids Res.*, 2007, **35**, D247–D252.
31 D. E. Harrison, R. Strong, Z. D. Sharp, J. F. Nelson, C. M. Astle, K. Flurkey, N. L. Nadon, J. E. Wilkinson, K. Frenkel, C. S. Carter, M. Pahor, M. A. Javors, E. Fernandez and R. A. Miller, *Nature*, 2009, **460**, 392–395.
32 S. Sarkar, B. Ravikumar, R. A. Floto and D. C. Rubinsztein, *Cell Death Differ.*, 2009, **16**, 46–56.
33 T. Pan, S. Kondo, W. Zhu, W. Xie, J. Jankovic and W. Le, *Neurobiol. Dis.*, 2008, **32**, 16–25.
34 C. Malagelada, Z. H. Jin, V. Jackson-Lewis, S. Przedborski and L. A. Greene, *J. Neurosci.*, 2010, **30**, 1166–1175.
35 T. R. Korfhagen, T. D. Le Cras, C. R. Davidson, S. M. Schmidt, M. Ikegami, J. A. Whitsett and W. D. Hardie, *Am. J. Respir. Cell Mol. Biol.*, 2009, **41**, 562–572.
36 S. M. Holland, F. R. DeLeo, H. Z. Elloumi, A. P. Hsu, G. Uzel, N. Brodsky, A. F. Freeman, A. Demidowich, J. Davis, M. L. Turner, V. L. Anderson, D. N. Darnell, P. A. Welch, D. B. Kuhns, D. M. Frucht, H. L. Malech, J. I. Gallin, S. D. Kobayashi, A. R. Whitney, J. M. Voyich, J. M. Musser, C. Woellner, A. A. Schäffer, J. M. Puck and B. Grimbacher, *N. Engl. J. Med.*, 2007, **357**, 1608–1619.
37 S. M. Kurian, R. Heilman, T. S. Mondala, A. Nakorchevsky, J. A. Hewel, D. Campbell, E. H. Robison, L. Wang, W. Lin, L. Gaber, K. Solez, H. Shidban, R. Mendez, R. L. Schaffer, J. S. Fisher, S. M. Flechner, S. R. Head, S. Horvath, J. R. Yates, C. L. Marsh and D. R. Salomon, *PLoS One*, 2009, **4**, e6212.
38 M. Krallinger, A. Valencia and L. Hirschman, *Genome Biology*, 2008, **9**(Suppl 2), S8.
39 J. Crim, R. McDonald and F. Pereira, *BMC Bioinf.*, 2005, **6**(Suppl 1), S13.
40 R. McDonald and F. Pereira, *BMC Bioinf.*, 2005, **6**(Suppl 1), S6.
41 D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan, *Neoplasia*, 2004, **6**, 1–6.
42 A. Bernthaler, I. Mühlberger, R. Fechete, P. Perco, A. Lukas and B. Mayer, *Mol. BioSyst.*, 2009, **5**, 1720–1731.
43 C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance, L. R. Sudarsky, D. G. Standaert, J. H. Growdon, R. V. Jensen and S. R. Gullans, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 955–60.
44 www.clinicalTrials.gov, 2010.